MODELING AND STRUCTURAL STUDIES OF SINGLE-STRANDED RNA VIRUSES

A Dissertation
Presented to
The Academic Faculty

by

Yingying Zeng

In Partial Fulfillment
Of the Requirements for the Degree
Doctor of Philosophy in Biology

Georgia Institute of Technology

May 2013

MODELING AND STRUCTURAL STUDIES OF SINGLE-STRANDED RNA VIRUSES

Approved By:

Dr. Stephen C. Harvey, Advisor
School of Biology
Georgia Institute of Technology

Dr. Christine E. Heitsch
School of Mathematics
Georgia Institute of Technology

Dr. Nicholas Hud
School of Chemistry and Biochemistry
Georgia Institute of Technology

Dr. Ingeborg Schmidt-Krey
School of Biology
Georgia Institute of Technology

Dr. Roger Wartell
School of Biology
Georgia Institute of Technology

Date Approved: March 15, 2013

Dedicated to My Grandfather

ACKNOWLEDGMENTS

I would like to thank my advisor Dr. Harvey for the insightful guidance on both my research and my future. I would never be able to complete my PhD studies without his support. I am also grateful to my committee members, Drs. Christine Heitsch, Nicholas Hud, Ingeborg Schmidt-Krey, and Roger Wartell for their advices and feedback over the years. I would also like to thank all the members of the Harvey lab, Anton, Jared, Shefaet, Bee, Shreyas, Ethan and Piyush for all the support, criticism and discussions. Finally, I thank my parents and my boyfriend Zhichao for their love and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

STMV	Satellite tobacco mosaic virus
HIV	Human immunodeficiency virus
MS2	Bacteriophage MS2
PaV	Pariacoto virus
STNV	Satellite tobacco necrosis virus
BPMV	Bean pod mottle virus
CCMV	Cowpea chlorotic mottle virus
T number	Triangulation number
SHAPE	Selective 2'-hydroxyl acylation and primer extansion
MLD	Maximum ladder distance
RNA	Ribonucleic acid
ssRNA	Single-stranded ribonucleic acid
PS	Packaging signal
cDNA	Complementary deoxyribonucleic acid
mRNA	Messenger RNA
tRNA	Transfer RNA
rRNA	Ribosomal RNA

SUMMARY

My research focuses on structures of the genomes of single-stranded RNA viruses. The work presented here covers the RNA sequence analysis, secondary structure studies, and the modeling of RNA tertiary structures. The first project is concerned with the sequence and secondary structure of HIV-1 RNA. Based on the secondary structure that Watts *et al*. predicted using SHAPE, I performed a series of analysis and the results suggested that a significant number of adenosines at the wobble position of the codons lead to an unusual structure with a large number of unpaired nucleotides. The findings indicated how the virus balances evolutionary pressures on the genomic RNA secondary structure against pressures on the sequence of the viral proteins.

The second project is the modeling of satellite tobacco mosaic virus (STMV). STMV is a T=1 icosahedral virus with a single piece of RNA that has 1058 nucleotides. X-ray crystallography studies of this RNA have revealed a structure containing 30 helices. The linkers between the helices, the possible structures at the interior of the icosahedron, and the sequence of the RNA were all absent in the crystal structure. To explore how the genome is organized within the protein capsid, I built a three-dimensional model based on the RNA secondary structure predicted by Susan Schroeder. Being the first all-atom model of any virus, this model is highly correlated with the crystal structure; and the comparison with the *in vitro* structure of the same RNA supports the hypothesis that the capsid protein plays an important role in RNA folding during assembly.

The third project includes the modeling of bacteriophage MS2 (MS2) and the examination of the compactness of RNA in different viruses. MS2 is a T=3 RNA virus,

and the cryo-EM studies have revealed a double-shell conformation of the genome. The final model of MS2 recaptures the double-shell structure of the RNA presented in the cryo-EM density. In addition, the predicted secondary structure that I used for the construction of the model shares a strong similarity with the *in vitro* structure determined in 1980s. This similarity contrasts with the striking difference between *in vivo* and *in vitro* RNA structures observed in STMV. Inspired by this finding, I examined the compactness of the RNA of several different viruses. The results strongly suggest that the RNAs of viruses requiring packaging signals have evolved to be structurally compact, which facilitates post-replicational RNA packaging. In contrast, viruses that do not depend on packaging signals probably adopt co-replicational RNA packaging.

CHAPTER 1

INTRODUCTION

Icosahedral ssRNA viruses

Icosahedral single-stranded RNA (ssRNA) viruses vary in size, from the small ones such as the 17 nm satellite tobacco mosaic virus (STMV) whose genome contains only 1058 nucleotides (1), to the larger ones like the picornaviruses (2), which has a RNA of 7-8 kb in length. Many pathogens belong to this class of viruses, including influenza A virus (3), hepatitis C virus (4), SARS coronavirus (5) and foot and mouth disease virus (6). Solving the structures of those viruses is essential for understanding the mechanisms of their assembly and hence developing drugs that may interfere with their reproduction.

An icosaheral ssRNA virus is composed of two parts: a protein capsid with icosahedral symmetry and one or more pieces of single-stranded RNA located inside the capsid (7). The protein capsid consists of copies of protein subunits. Those subunits gather into pentamers and hexamers, which together form an icosahedral structure. Each viral capsid is assigned a triangulation number (T number); and a virus with the triangulation number of T has 60T protein subunits assembling into 12 pentamers and 10(T-1) hexamers (Figure 1.1) (8,9). Taking advantage of the icosahedral symmetry of the protein capsid, scientists have solved the capsid structures of many viruses by using icosahedral averaging during 3D-reconstruction from X-ray crystallography.

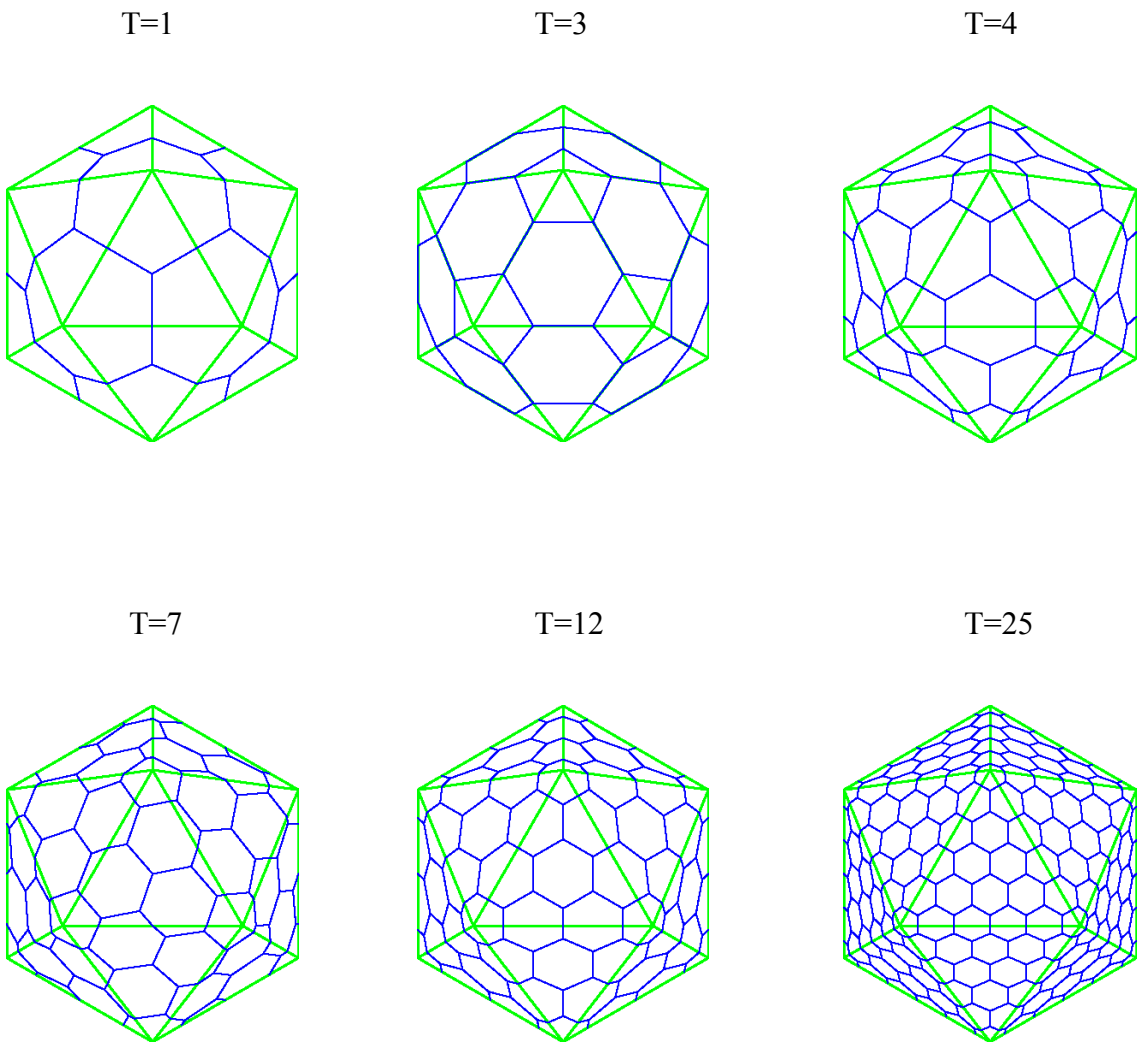

Figure 1.1: Protein capsids of different T numbers. Images created at http://viperdb.scripps.edu/

Compared with the protein capsid, the structure of the genome is more difficult to determine, since the RNA inside the capsid does not possess true icosahedral symmetry. X-ray crystallography on several viruses has revealed double-helical structures of RNA underneath the capsid, suggesting that the genome of an icosahedral ssRNA virus folds into a geometric pattern that presents certain level of icosahedral symmetry (Figure 1.2) (10). Larson *et al.* discovered in STMV a helical structure of nine base pairs along each

of the 2-fold axes (11); Tang *et al.* found that the genome of pariacoto virus (PaV) forms a dodecahedral cage of duplex RNA (12); and cryo-EM studies on bacteriophage MS2, together with x-ray crystallography on MS2 viral-like particles, also suggested that the RNA forms stem-loop structures (13-15). Those double-helical regions of RNA interact with the protein capsid and play an important role in viral assembly (16,17). However, except for the RNA double-helical regions that form an icosahedrally symmetrical pattern, the other parts of the RNA are invisible in the crystal structure or cryo-EM density due to the lack of symmetry. In addition, the RNA sequence information is also absent in the crystal structure because of the icosahedral averaging. The lack of information about the organization of the entire RNA drives modelers to predict the detailed RNA structures within the viral particles using computational methods.

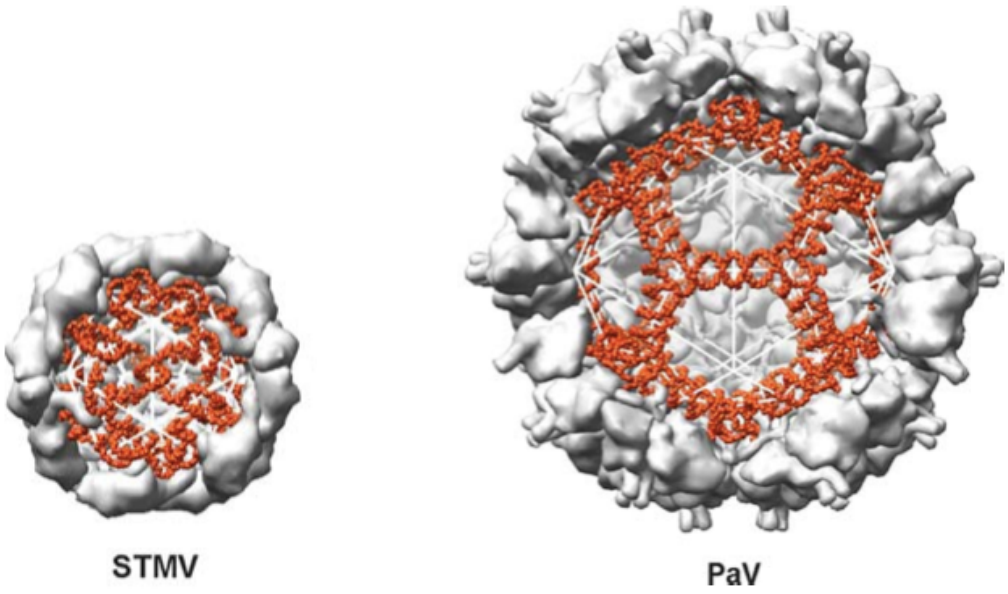

Figure 1.2: RNA double-helices observed in crystal structures of STMV and PaV (Schneemann *et al.,* 2006 (10)).

Structural virology

The study of viral structures is essential to understand the mechanism of viral assembly and infection. Different methods have been developed to solve the structures of viruses, and the most frequently used ones are X-ray crystallography and cryo-electron microscopy (cryo-EM). Both methods take advantage of the symmetry of a viral capsid, obtaining an averaged structure or electron density. Besides experimental approaches, molecular modeling provides insights into the structural details, especially for the regions that are not visible in the crystal structure or cryo-EM density.

X-ray crystallography

X-ray crystallography is a power tool that is used to solve the structures of viruses. It produces high-resolution models referred to as the "crystal structures" of the molecules. In the process of virus crystallography, homogenous and purified viral particles are crystallized and the X-ray diffraction data are collected (18). A crystal structure of atomic resolution can be obtained after data processing and model refinement. This method has been improved over the years and become relatively routine to solve structures of viruses with various sizes (18).

The first virus to be crystallized was the tobacco mosaic virus (TMV) (19). X-ray crystallography studies on TMV revealed the helical form of this virus: the protein subunits gather into a helix with 16.3 subunits per turn (Figure 1.3); the single-stranded RNA, located 40 Å from the central axis, also adopts a helical form and is intercalated between the turns of the protein helix (20,21). The resolution of the maps was improved over the years from 12 Å to 2.8 Å, which provides information about the structural details

(20). It also provides insights into the assembly process, indicating that the insertion of a RNA stem-loop into the central hole of the protein triggers the dislocation of the protein disk into a helix (20).

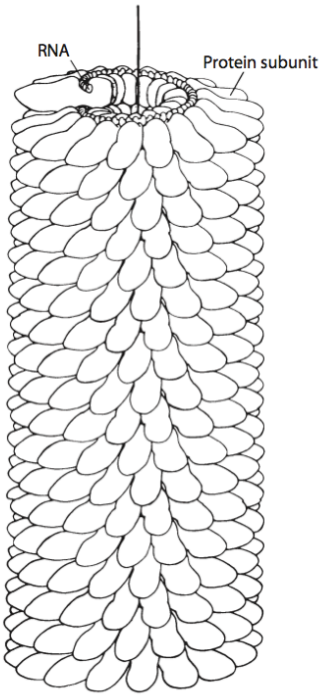

Figure 1.3: Diagram of TMV structure (Klug, 2010 (21)).

The structures of many viral capsids have been solved so far using X-ray crystallography, especially the ones that have icosahedral symmetry. For example, the crystal structure of the capsid protein of southen bean mosaic virus was solved at 2.8 Å (22). In addition to capsids, X-ray crystallography studies have also led to discoveries of protein-RNA interactions in many viruses, such as bean pod mottle virus, whose trefoil-shaped RNA binds to the side chains of the protein subunit (23). Similar interactions were also found in flock house virus (24) and Pariacoto virus, where the RNA duplexes interact with the

N-termini of the protein subunits (Figure 1.4) (12). Obviously, X-ray crystallography has become a standard approach that is essential in structure determination of viral particles.

Figure 1.4: Interactions between protein N-termini and RNA helix in Pariacoto virus (Tang *et al.*, 2001 (12)).

Cryo-electron microscopy (Cryo-EM)

With the developments in electron microscopy, cryo-EM has become an increasingly powerful tool in structural virology, and it provides information that complements crystallography data (25). X-ray crystallography has limitations, for example, a large, complicated or labile molecule may not be able to pack into a crystal (26). Also, crystallization fails when the sample is not pure (25,26). Although cryo-EM, in most cases, cannot achieve as a high resolution as crystallography, it has several advantages. First, it does not have specific requirements on sample purity. Second, it can be applied to large molecular complexes. In addition, the cryo-EM sample is rapidly frozen to the

temperature of liquid nitrogen, which transforms the water into vitreous ice (25,26). In this way, the sample is maintained in its native state, facilitating the study of dynamics of macromolecules.

The combination of cryo-EM and X-ray crystallography has become a standard approach for studying viral structures. Atomic-level structures can be obtained by fitting the crystal structures into the cryo-EM density. The combination of the two methods also enables the investigation of conformational changes during viral assembly. For example, cryo-EM studies on HK-97, together with X-ray crystallography data, revealed that the protein capsid expands 25% during the viral maturation (27,28). Additionally, cryo-EM is able to capture physiological states such as viral attachment to the cell. An example is the 5-rotational averaged image obtained from bacteriophage MS2 attaching to the F-pilus of *E.coli* (Figure 1.5) (29). With the improved technology, single-particle analysis from cryo-EM, in certain cases, can achieve a resolution of 4 Å.

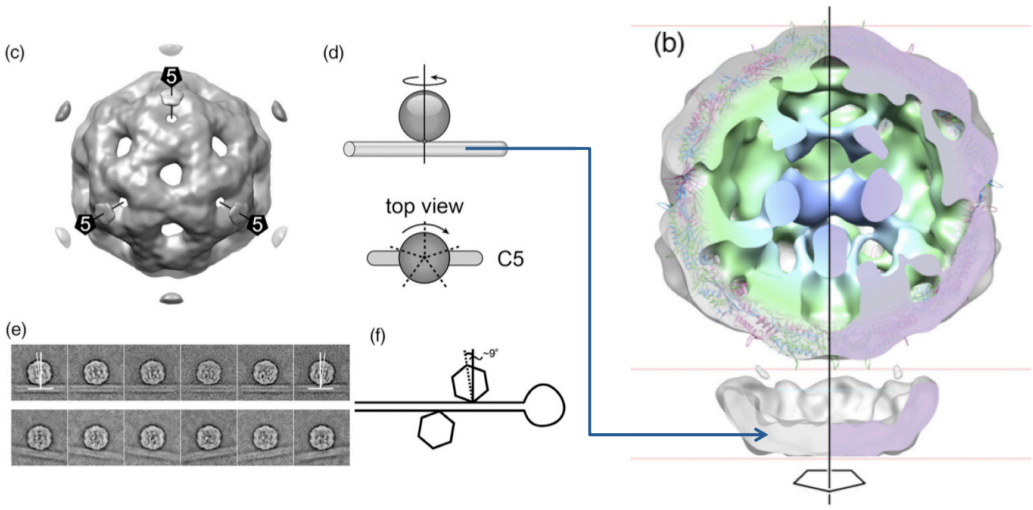

Figure 1.5: Cryo-EM density of bacteriophage MS2 obtained from 5-fold rotational averaging (Toropova *et al.*, 2011 (29))

Modeling of small icosahedral ssRNA viruses

As mentioned earlier, x-ray crystallography and cryo-EM only reveal parts of the structures of the genomes in icosahedral ssRNA viruses. For example, the visible nucleotides of a STMV particle account for only 57% of the entire RNA (11), and only 35% of the PaV genome was visible in the crystal structure (12). As a result, molecular modeling of icosahedral ssRNA viruses became important in providing insights into the detailed RNA structures and the organizations of the genomes in three-dimensional space. So far modeling has been done on several small viruses. Freddolino *et al.* constructed an all-atom model for STMV (30) based on the hypothesis that STMV RNA folds into a structure of 30 stem-loops connected by single-stranded regions (17) (Figure 1.6). The RNA in their model has the same size as that of the wild-type viral RNA, but with an artificial sequence because of the lack of sequence information in the crystal structure. Through molecular dynamics simulation, they found that RNA plays an important part in stabilizing the viral particle (30).

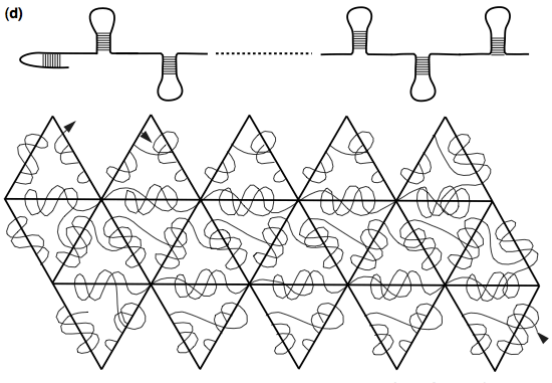

Figure 1.6: A model of STMV RNA secondary structure and the arrangement of the stem-loops to form an icosahedron (Larson *et al.*, 2001 (17)).

Another all-atom model based on a crystal structure is the PaV model constructed by Devkota *et al.* (31). The final model included all the missing parts in the crystal structure and had a similar radial density distribution as the crystal structure. As in the STMV model built by Freddolino *et al.,* the RNA in this model did not have the real sequence. The helices forming the dodecahedron are directly from the crystal structure, and the rest of the genome was represented by twelve copies of a domain in 23S rRNA. Unlike the STMV model where the stem-loops occupy the edges of the icasahedron, the PaV model has the stem-loops in the core while the long-range helices along the dodecahedral edges (Figure 1.7). Neither of the STMV nor PaV model is based on a plausible secondary structure that is experimentally or computationally predicted. Therefore, I would like to explore how a secondary structure of the real sequence can be arranged into an icosahedral pattern by developing new methods of viral RNA modeling.

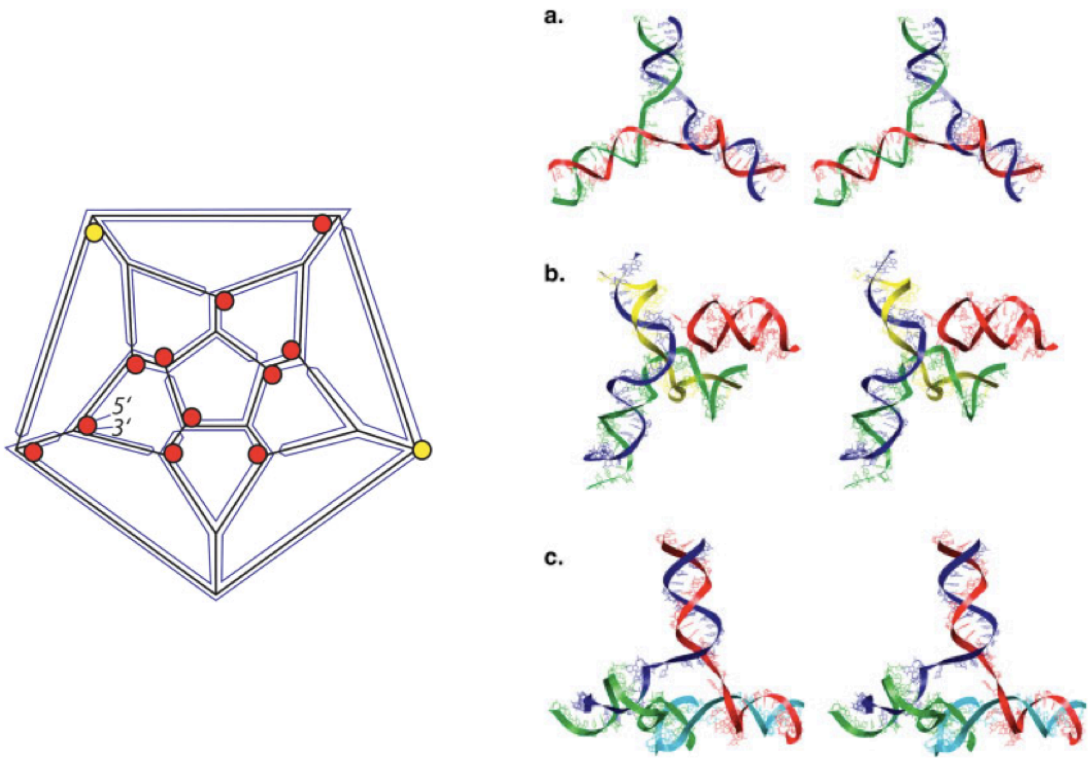

Figure 1.7: RNA secondary structure in PaV model. Left: secondary structure map. Red and yellow circles represent three- and four-way junctions respectively. Right: stereo images of junctions at the vertices. (a) Three-way junction (b) Three-way junction with a "stalactite" dropping into the interior of the virus (c) Four-way junction (Devkota *et al.*, 2009 (31))

RNA secondary structure prediction

RNA secondary structure refers to the base pairing interactions of an RNA molecule (Figure 1.8). The secondary structure of a RNA provides valuable information on its tertiary conformation, as it indicates the locations of the helices within the nucleotide chain. Various methods have been developed to predict RNA secondary structures, including both experimental and computational approaches. Some of the most frequently used prediction methods are thermodynamic programs and chemical probing.

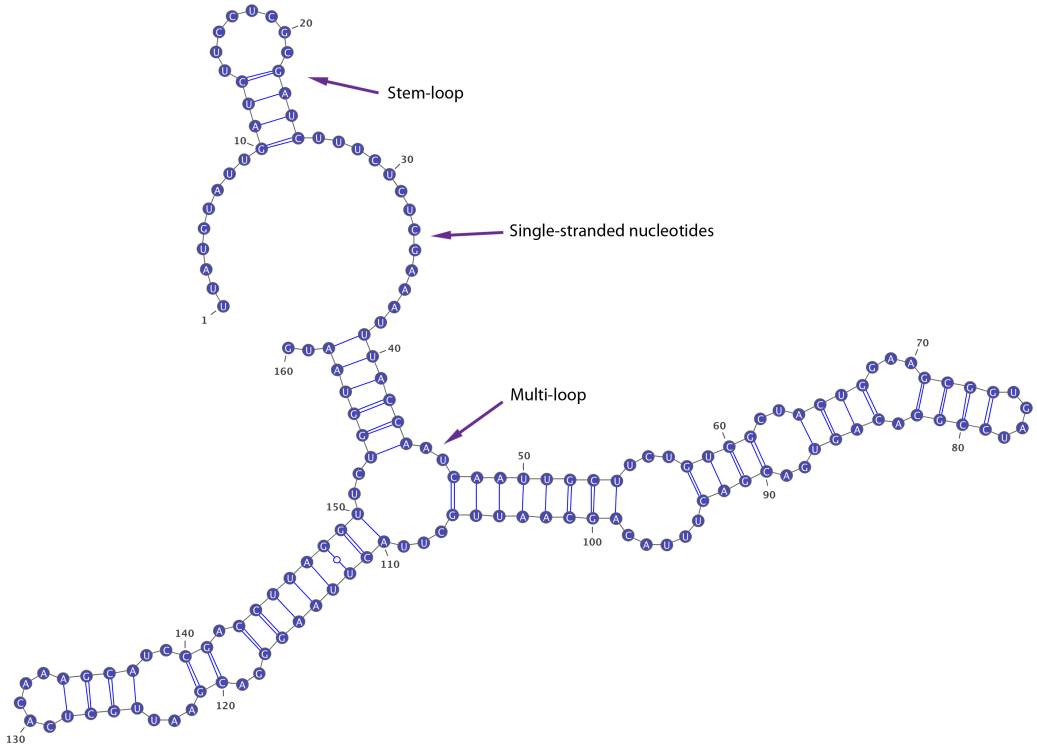

Figure 1.8: RNA secondary structure

Thermodynamic prediction programs

Thermodynamic prediction programs, such as RNAfold (32,33), UNAFold (34) and GTfold (35), predict RNA secondary structures from sequences. The calculation of the energy relies on the nearest neighbor model (36), in which the energy of a base pair depends on the properties of the adjacent base pairs. An energy dot plot (Figure 1.9), which stores the energy of each possible base pair, is generated during the prediction. There are two algorithms for generating a secondary structure. Minimum free energy (MFE) algorithms calculate the lowest free energy and suboptimal structures through a traceback process (37); partition function algorithms calculate the base pairing

probability for each nucleotide and generate an ensemble of structures that are drawn with probabilities based on their Boltzmann weights (38).

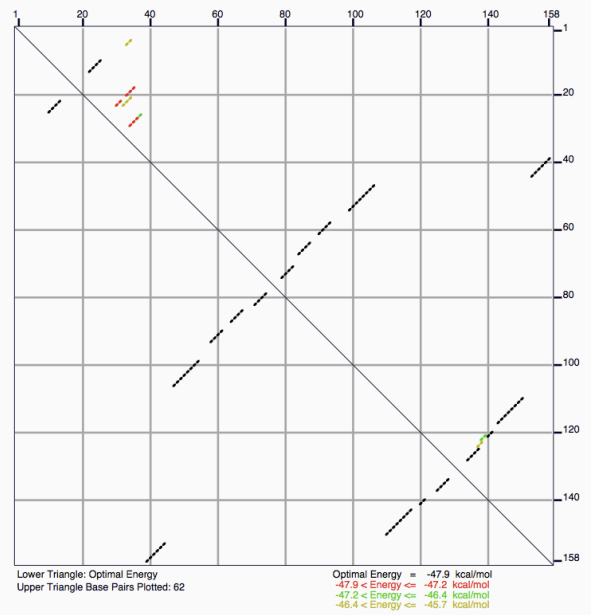

Figure 1.9: An energy dot plot

The RNA energy parameters used in those programs are derived from Turner group (37). Due to the limitations of the experiments and lack of knowledge in all the details about RNA folding, the energy parameters are not always accurate, especially for the multi-loop regions. Therefore those thermodynamic programs, in most cases, fail to select the correct structure when predicting the secondary structure for a known RNA. Nevertheless, thermodynamic predictions provide information regarding the energy range and the possible base-pairing pattern of a RNA molecule, and they are able to rapidly generate a large number of structures, which enables researchers to evaluate multiple RNA molecules at a time.

SHAPE

SHAPE stands for selective 2'-hydroxyl acylation and primer extansion. It is a secondary structure probing method that combines chemical probing and thermodynamic prediction (39). The main steps of SHAPE include RNA modification, primer extension, capillary electrophoresis and data processing. During the experiment, a chemical reagent such as NMIA is used to modify the 2'-hydroxyl group the nucleotides. The 2'-hydroxyl group of a single-stranded nucleotide is more flexible than that of a double-stranded nucleotide, because the 2'-hydroxyl group in a paired nucleotide is constrained in the RNA C3'-endo conformation. As a result, the single-stranded nucleotides are more likely to be modified by the reagent (Figure 1.10). After the modification, the primer extension process produces cDNA with different lengths that correspond to the locations of the modifications (Figure 1.11). The lengths of cDNA are detected using capillary electrophoresis. The signals are then converted to SHAPE reactivities using the software ShapeFinder (Figure 1.12) (40). The reactivity of each nucleotide is incorporated into RNAstructure (41), a thermodynamic prediction program, as a pseudo-energy ΔG_{SHAPE} (Eq. 1). The slope m and intercept b in Eq. 1 are parameterized against the 23S rRNA structure that was determined using comparative sequence analysis. The optimized values for m and b are selected from predictions that have high sensitivity and positive predicted value (PPV) (sensitivity: the number of correctly predicted base pairs divided by the total number of base pairs in the known structure; PPV: the number of correctly predicted base pairs divided by the total number of base pairs in the predicted structure). Tests on several RNA molecules demonstrated that SHAPE provides satisfactory accuracy (Table

1.1). As a result, it has been extensively used on various RNAs, including the entire genome of HIV-1 RNA (42).

$$\Delta G_{SHAPE}(i) = m \bullet \ln\left[SHAPEreactivity(i) + 1\right] + b \qquad \text{Eq. 1}$$

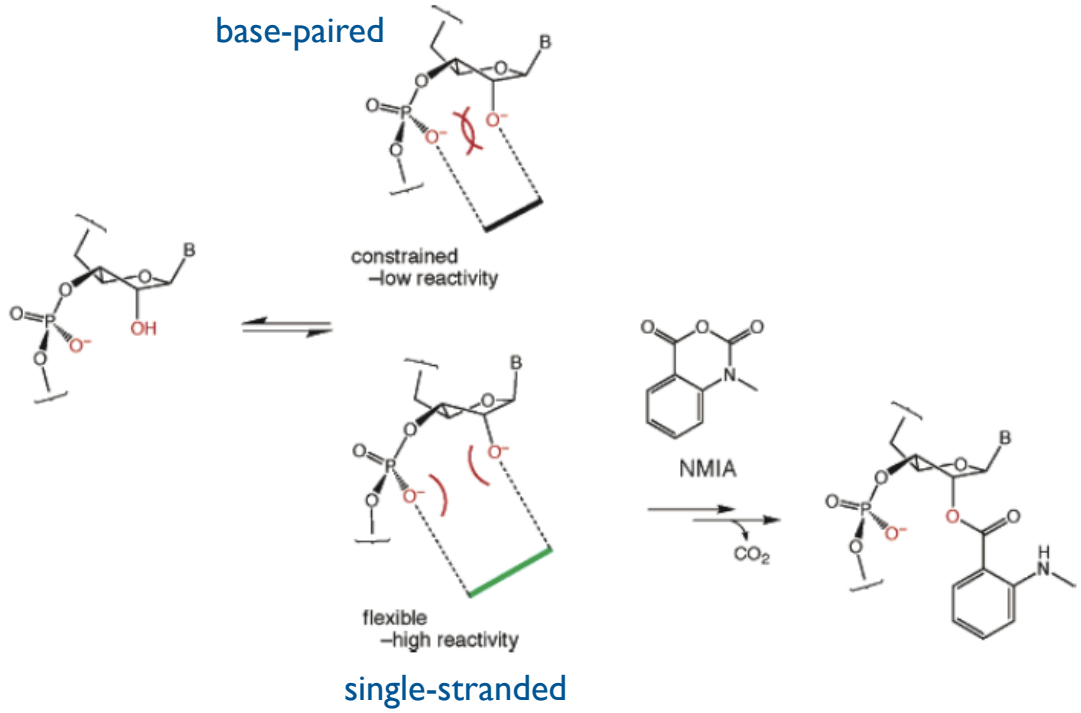

Figure 1.10: NMIA modification of single-stranded nucleotide (Deigan *et al.*, 2009 (39))

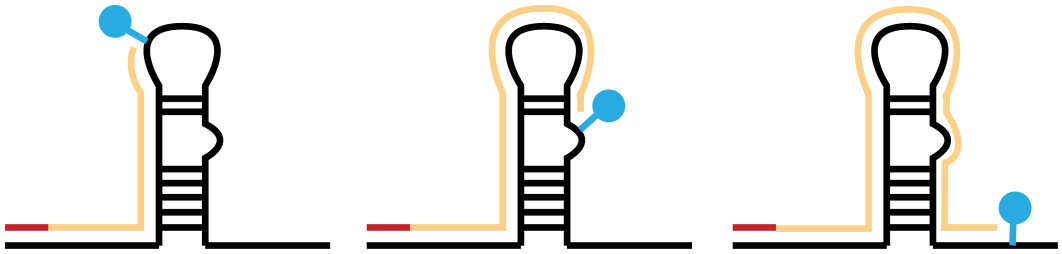

Figure 1.11: Primer extension. (39)

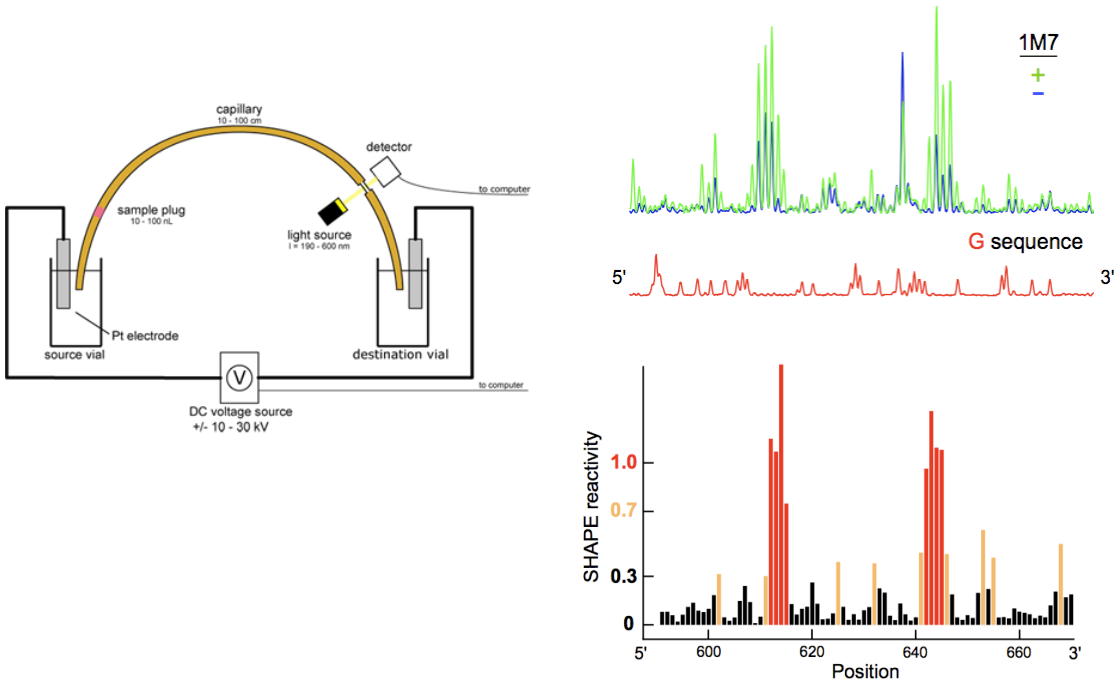

Figure 1.12: Capillary electrophoresis (left) and signal processing (right). (Deigan *et al.*, 2009 (39))

Table 1.1: Prediction accuracies for various RNAs (Deigan *et al.*, 2009 (39))

		No constraints		SHAPE	
RNA	Nucleotides	Sensitivity	PPV	Sensitivity	PPV
Yeast $tRNA^{Asp}$	75	95.2	95.2	100.0	100.0
HCV IRES domain II	95	56.5	59.1	95.7	100.0
P546 domain, group I intron	155	42.9	44.4	96.4	98.2

MLD

Maximum ladder distance (MLD) was first introduced by Yoffe *et al.* to measure the extendedness of a RNA secondary structure (43). Ladder distance between nucleotide i and j is defined as the number of base pairs that are crossed along the most direct path from base i to base j in the two-dimensional secondary structure graph, and the MLD measures the longest direct path across a secondary structure (Figure 1.13). MLD is considered to represent the contour length of the RNA in three dimensions, and hence it reflects the radius of gyration of the molecule. Yoffe *et al.* used MLD to examine the viral RNA secondary structures. They calculated the averaged MLD for a predicted ensemble of secondary structures for both viral RNA and random sequences, and their results suggested that viral RNA have been evolved to form more compact structures than random sequences, because of the selection pressure produced by the small volume enclosed by the capsid. Athavale *et al.* (44) also used this method to study STMV RNA, and the results indicated that STMV RNA is more extended than random sequences, which contradicts with Yoffe *et al*'s findings. The uniqueness of STMV RNA will be further discussed in chapter 3 and 4.

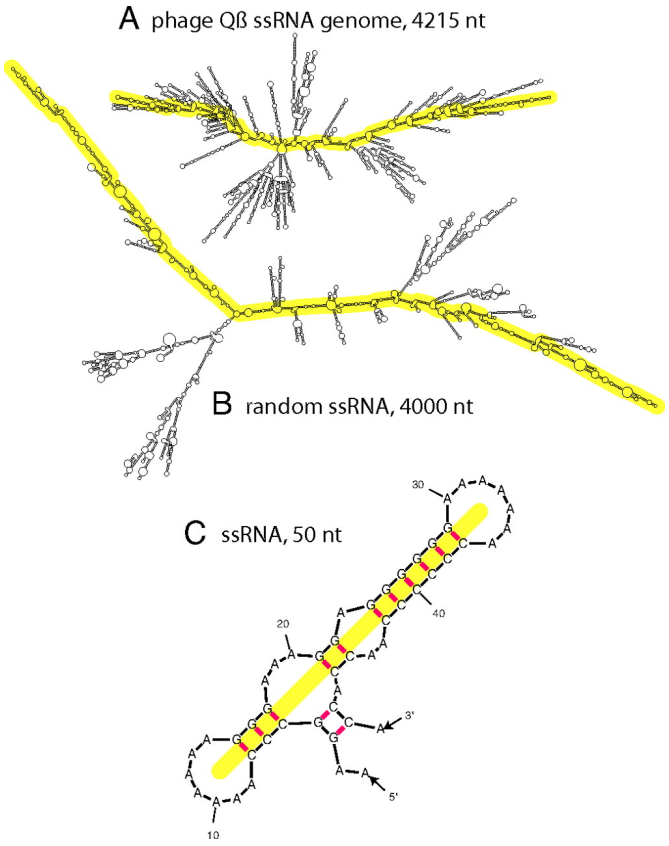

Figure 1.13: MLD of RNA molecules (Yoffe *et al.*, 2008 (43))

CHAPTER 2

SEQUENCE AND SECONDARY STRUCTURE ANALYSIS OF HIV-1 RNA

Abstract

The secondary structure of the RNA genome of HIV-1, determined by SHAPE, is very unusual. Only 31% of the 9142 nucleotides are involved in either Watson-Crick or wobble base pairs. This is only about half the frequency of base pairing that is found in ribosomal RNAs, and in secondary structures predicted for RNAs containing random sequences. Watts *et al.* showed that the secondary structure plays several functional roles, including translational regulation, protein folding, and splicing. I have examined the composition and sequence of the HIV-1 genome, in an effort to understand the origins of its unusual secondary structure. The RNA is very A-rich (36%) and C-poor (18%). These disparities are not uniformly distributed throughout the genome: the collection of non-base-paired nucleotides is even more A-rich and C-poor than the genome as a whole. In the coding regions of the genome, compositional disparities are concentrated in the wobble position of the codons. This is evidently how the virus balances evolutionary pressures on the genomic RNA secondary structure against pressures on the sequences of the viral proteins.

Introduction

Human immunodefiency virus (HIV) is the causative agent of acquired immunodeficiency syndrome (AIDS). Conserved structural elements in the single-stranded RNA genome serve important regulatory functions in infection and replication (42), so the identification of these elements has important implications for the regulation of gene expression by secondary structures in other mRNAs, particularly the genomes of other RNA viruses.

There are a number of methods for predicting RNA secondary structure. Comparative sequence analysis (45) examines multiple sequences for a given molecule and predicts base pairing based on covariation analysis; this is how the secondary structures of ribosomal RNAs were determined (46,47). Thermodynamic predictions for a single sequence can be generated by a variety of programs, such as Mfold (48,49), UNAfold (34), RNAfold and RNAstructure (50). Finally, RNA folding occurs during transcription, so that some elements of secondary structure may be kinetically trapped; this has led to efforts to include kinetic effects into thermodynamic methods (51).

Kevin Weeks and his collaborators have shown (39) that RNA secondary structure can be determined quite accurately for large RNAs by using a weighted combination of thermodynamic analysis and data from a high-throughput chemical probing technique, SHAPE (selective 2'-hydroxyl acylation by primer extension) (52). The SHAPE reactivity of a given nucleotide is correlated with the probability that the base is unpaired (53). RNA secondary structure can then be established by supplementing information from traditional thermodynamic predictions with properly weighted information from

SHAPE reactivity. The idea of combining other information with thermodynamic predictions is not new (54), but the specificity of SHAPE probing is apparently higher than that of other reagents. The ShapeFinder program (40) analyzes the experimental data to determine the reactivity of each nucleotide, and RNAstructure (41) then combines suitably weighted reactivities with thermodynamic predictions to generate the secondary structure. The current weights were determined by analysis of a ribosomal 23S RNA and the accuracy of the approach was demonstrated by comparing the predicted secondary structure of 16S rRNA with the known structure (39).

Watts *et al.* have determined the secondary structure of the entire 9173-nucleotide genome of HIV-1 by SHAPE analysis. There are at least ten structured regions, and there are six large regions (~200-600 nt each) that are essentially unstructured (Figure 2.1). The latter is a novel observation, revealing that the secondary structure of the HIV genome is unlike that of any previously analyzed RNA. Overall, the fraction of unpaired bases is much higher (59%) than is characteristic of structural RNAs such as the 16S and 23S rRNAs, or of random sequence RNAs containing equimolar fractions of A, C, G and U. This clearly facilitates translation, but Watts *et al.* also found evidence that the secondary structure plays a range of regulatory roles. For example, regions with high levels of secondary structure in the mRNA correspond to boundaries between protein domains, while splice site acceptors and hypervariable regions are largely unstructured.

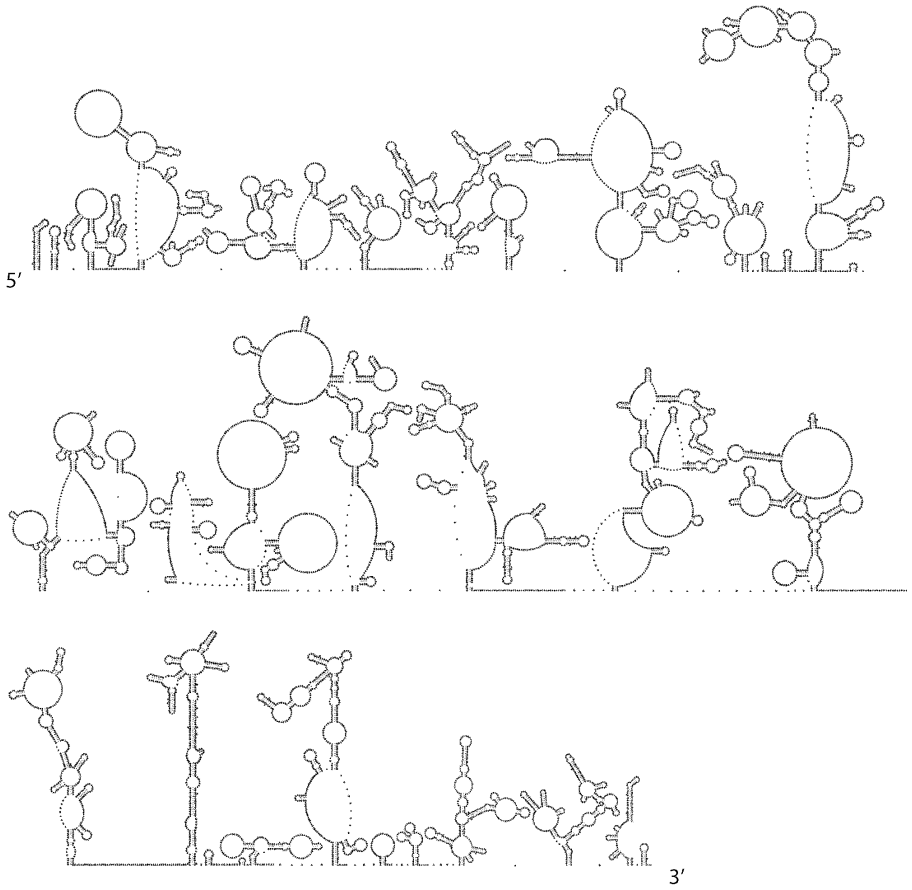

Figure 2.1: HIV-1 RNA secondary structure determined by Watt *et al.* (42)

This result raises a series of questions. What unique characteristics of the HIV-1 genome are responsible for the low level of secondary structure? What role does composition play? Does the overall composition have any remarkable features? Is the composition of the large single-stranded regions similar to the overall composition, or is it unusual in any way? Beyond composition, does sequence contribute to the propensity for single-strandedness? That is, for a region of given composition, is the extent of secondary structure significantly lower than would be predicted for a random sequence of the same

composition? As will be seen, both the composition and sequence do, indeed, have unusual properties. The genome is a messenger RNA, which must place some constraints on composition and sequence, so I tested the hypothesis that the viral sequence exploits the flexibility of the wobble position to generate the largely unstructured regions. Finally, I examined over 1000 HIV genomes, to determine if the composition and sequence effects associated with the NL4-3 sequence used in the SHAPE study (42) are common to all strains.

In addition, since the reported HIV secondary structure has so many unpaired nucleotides, I wondered if there might be some pieces of secondary structure that were missed in the SHAPE analysis. The ShapeFinder weights were derived from a ribosomal RNA, which has far fewer unpaired nucleotides than the HIV-1 model published by Watts *et al.* (42), and it is possible that those weights are not appropriate for largely unstructured RNAs. To test this, I analyzed the large putatively unpaired regions using purely thermodynamic predictions, to see if I could find any likely secondary structures that were not reported by Watts *et al*.

Methods

Chi-square test

The Chi-square test was used to evaluate whether the observed composition of a sequence is significantly different from its expected composition (the composition of a reference sequence). The formula for the chi-square test is:

$$\chi^2 = \sum_i \frac{(f_i - F_i)^2}{F_i}$$

For a given sequence, f_i is the observed count of one of the four nucleotides (A, G, C, U), and F_i, which is calculated using the composition of the reference sequence, is the expected count of that particular nucleotide.

RNAfold

RNAfold is a program from Vienna Package (33) (http://www.tbi.univie.ac.at/RNA/). It predicts RNA secondary structure using thermodynamic parameters. The secondary structures and base pairing probabilities in this analysis were predicted and calculated using the partition function command in RNAfold. Base pairing probability for each nucleotide was calculated by summing the probability of every possible base pair formed by that nucleotide.

UNAFold

UNAFold (34), developed by Zuker's lab, is also a thermodynamic prediction program for RNA and DNA. Some of the parameters that it uses are different from RNAfold. It is used in this analysis as a control method so that I can compare the results from different prediction programs.

Z-score

Z-score measures how many standard deviations an observation is above or below the mean. The formula for Z-score is:

$$Z = \frac{\Delta G^* - \langle \Delta G \rangle}{\sigma}$$

where ΔG^* is the free energy of the actual secondary structure; $\langle \Delta G \rangle$ is the mean of the free energies of the secondary structures formed by random sequences; and σ is the standard deviation of the population.

Spearman-correlation

Spearman's rank correlation coefficient is a non-parametric measure of correlation. It is a form of the Pearson coefficient with the data converted to rankings.

The Spearman coefficient is denoted with the Greek letter rho (ρ).

$$\rho = 1 - \frac{6 \times \sum d_i^2}{n \times \left(n^2 - 1\right)}$$

where d_i is the difference between the ranks of each observation on the two variables and n is the number of values in each data set.

RNA structure display

The RNA secondary structures were rendered using XRNA (http://rna.ucsc.edu/rnacenter/xrna/xrna.html).

Results and Discussion

The abundance of adenosines results in a largely unpaired structure of HIV-1 RNA

The secondary structure determined by Watts *et al.* shows a remarkably small fraction (41%) of nucleotides that form base pairs (Figure 2.1). By comparison, 61% of the residues in the *E. coli* 16 S rRNA are base paired (58% in the 23S rRNA). To explore the reasons behind such a low frequency of base pairs, I examined the composition of HIV-1 RNA. The results indicate that HIV-1 RNA is particularly enriched in adenosine residues (36%, *vs.* 25% and 26% for the 16S and 23S rRNAs) (Table 2.1). It was previously observed that the ribosomal RNA from the *Bos Taurus* mitochondrion has a very low fraction of base paired nucleotides (45%), and that it is very rich in adenosines (38%). Apparently, highly skewed compositions are correlated with low overall base pairing probabilities, at least for large RNAs.

Table 2.1: The composition of the entire HIV RNA sequence.

	Number of nucleotides	Percentage
A	3283	35.79%
G	2214	24.14%
C	1635	17.82%
U	2041	22.25%

While the overall composition is striking, this alone cannot account for the high degree of single-strandedness in the HIV genome. Upon closer inspection, I found that the extremes of composition in the HIV-1 RNA are concentrated in the single-stranded regions of the secondary structure. (Throughout the present work, I use "single-stranded" to denote any nucleotide that does not form a Watson-Crick or wobble base pair.) Of the 5391 residues that comprise these regions, 48% are adenosines (*vs.* 36% for the entire genome). A chi-square (χ^2) test rejects the null hypothesis that the single-stranded regions have the same composition as the total genome, at a confidence level $P < 10^{-88}$ (Table 2.2); adenosine is, of course, the largest contributor to χ^2. These results correlate with the previous observation that single-stranded regions of ribosomal RNAs also have skewed compositions, being more purine-rich than the entire rRNA (18); that same study also reported purine enrichment in the single-stranded regions of thermodynamically predicted secondary structures for random RNA sequences.

Table 2.2: Chi-square analysis of the entire single-stranded subsequence. ($\chi^2_{0.05} = 7.82$)

	Observed	Expected	$\left(f_i - F_i\right)^2 / F_i$
A	2597	1929.98	230.53
G	1009	1299.23	64.83
C	629	959.60	113.90
U	1156	1202.19	1.77
			$\chi^2 = 411.04$

Analyzing the secondary structure domain by domain shows that composition is similarly skewed on a local scale. I have broken the HIV-1 genome into 24 domains (Figure 2.2,

Table 2.3), and Figure 2.3 shows that, in most cases, the single-stranded regions are even richer in adenosine residues than is the domain as a whole. Table 2.3 defines the domains and presents the results of chi-square tests on each one, with the null hypothesis that the composition in the single-stranded regions is identical to that of the whole domain. The null hypothesis is rejected with $P < 0.05$ in 19 out of 24 tests.

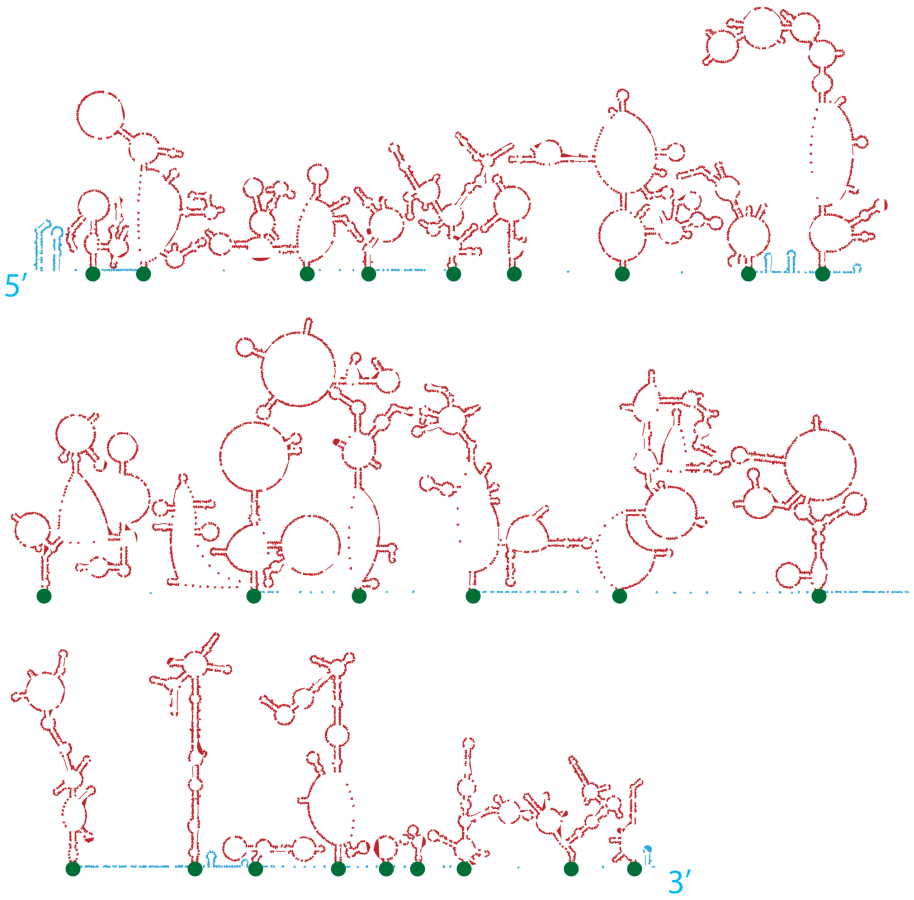

Figure 2.2: 24 domains of HIV-1 RNA. Each green dot marks the end of a domain.

Table 2.3: HIV-1 RNA domains and chi-square values for the hypothesis that single-stranded regions in the domains have the same composition as the whole sequence. ($\chi^2_{0.05}$=7.82) Numbers in red indicate statistical significance.

Domain	Region	χ^2	Domain	Region	χ^2
1	106-343	**8.8**	13	5139-5675	**14.9**
2	363-750	**19.9**	14	5725-6314	**26.7**
3	752-1172	**23.5**	15	6328-6798	**12.2**
4	1177-1312	5.37	16	6839-7188	**15.2**
5	1341-1795	**40.2**	17	7245-7599	**17.5**
6	1796-1946	**17.0**	18	7638-7778	7.1
7	1948-2545	**17.7**	19	7792-8218	**22.4**
8	2547-2778	**15.0**	20	8226-8268	1.1
9	2846-3381	**32.7**	21	8275-8348	7.0
10	3404-3943	**16.3**	22	8358-8684	**17.6**
11	3945-4518	**23.6**	23	8686-9009	**15.3**
12	4539-5135	**22.4**	24	9011-9139	7.3

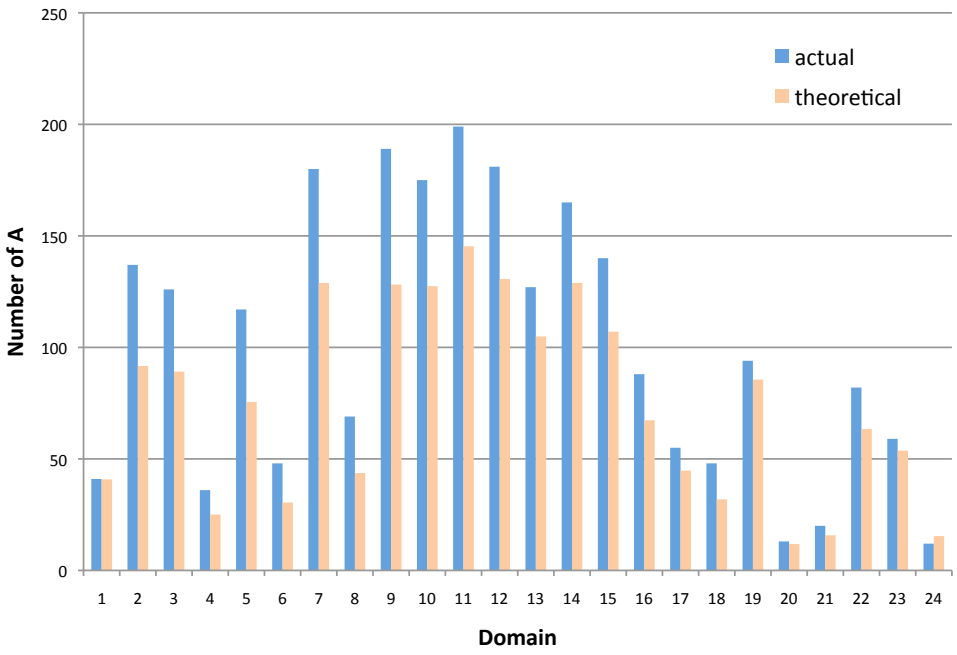

Figure 2.3: The comparison between the actual and theoretical numbers of A in 24 domains. The theoretical numbers are calculated from the composition of the entire HIV-1 RNA sequence.

Sequence effects also contribute to low base pairing frequencies

To test the idea that, for a region of a given composition, singled-strandedness is also promoted by sequence effects, I generated an ensemble of model RNAs in which the real sequence (and base pairing) is preserved in the double-stranded regions, but in which the sequence in the single-stranded regions is shuffled, with a different shuffle for each model. Each member of the ensemble has the true composition in the single-stranded regions, and the real secondary structure can be used as a constraint for secondary structure prediction by thermodynamic methods (Figure 2.4). I then asked whether the number of base pairs that the prediction adds to the published structure is the same for the native sequence as it is for the shuffled sequences.

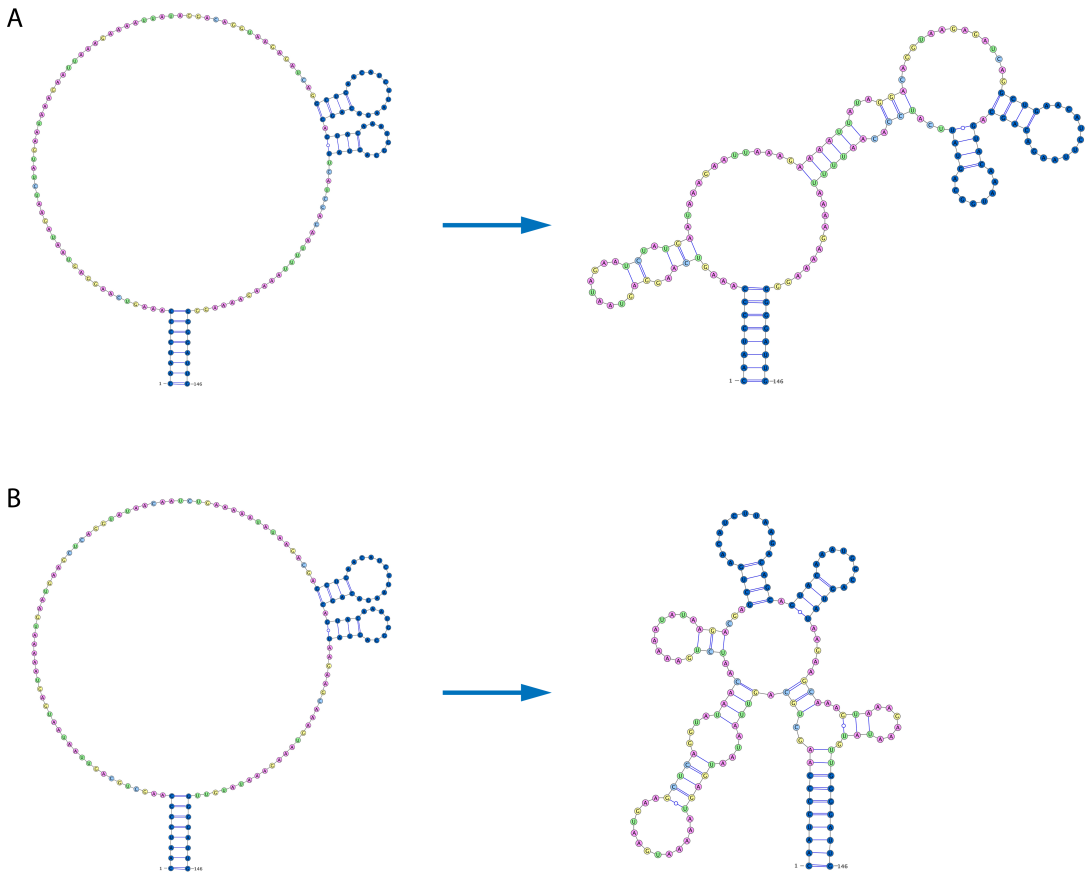

Figure 2.4: An example of the actual (A) and shuffled (B) sequences. The dark blue regions are helices/stem-loops in the published structure. Their sequences remain unchanged during the shuffling and their structures are constrained during the prediction. The images on the left side of the arrows indicate the input sequence with structural constraints, and the images on the right side of the arrows are the output structures from RNAfold.

If the high degree of single-strandedness in the HIV-1 genome is due only to the skewed composition in single-stranded regions, then the predicted secondary structure of the actual sequence should have about the same extent of base pairing as does the ensemble of model RNAs. To provide a quantitative test of this idea, I generated 1000 model RNAs

for each domain (1000 shuffled single-stranded regions) and predicted the secondary structures of all of them. For a given domain, the free energy distribution for the ensemble of folds is described by some mean value $<\Delta G>$ and some standard deviation s. If the stabilization free energy of the actual secondary structure is ΔG^*, then I can characterize it relative to the distribution of all model folds using a Z score that measures how many standard deviations ΔG^* is above or below the mean of the distribution,

$$Z = (\Delta G^* - <\Delta G>)/s$$

If sequence within the single-stranded regions has no effect, then Z has an expected value of zero with standard deviation s. If the actual secondary structure has more base pairing than a typical model secondary structure, Z will be negative, while a positive value of Z will indicate that the actual secondary structure has less base pairing than expected. Figure 2.5 compares the predicted minimum free energy (MFE) upon folding the actual sequence with the average MFE obtained by folding the shuffled ensemble. It shows that in 23 of the 24 domains, the actual secondary structure has less base pairing than a typical model secondary structure. Table 2.4 quantifies these results with the corresponding Z scores, using MFE and centroid structures predicted by RNAfold. Clearly, single-strandedness is not simply due to composition: the sequences of single-stranded regions of the HIV-1 genome are organized to hinder base pairing.

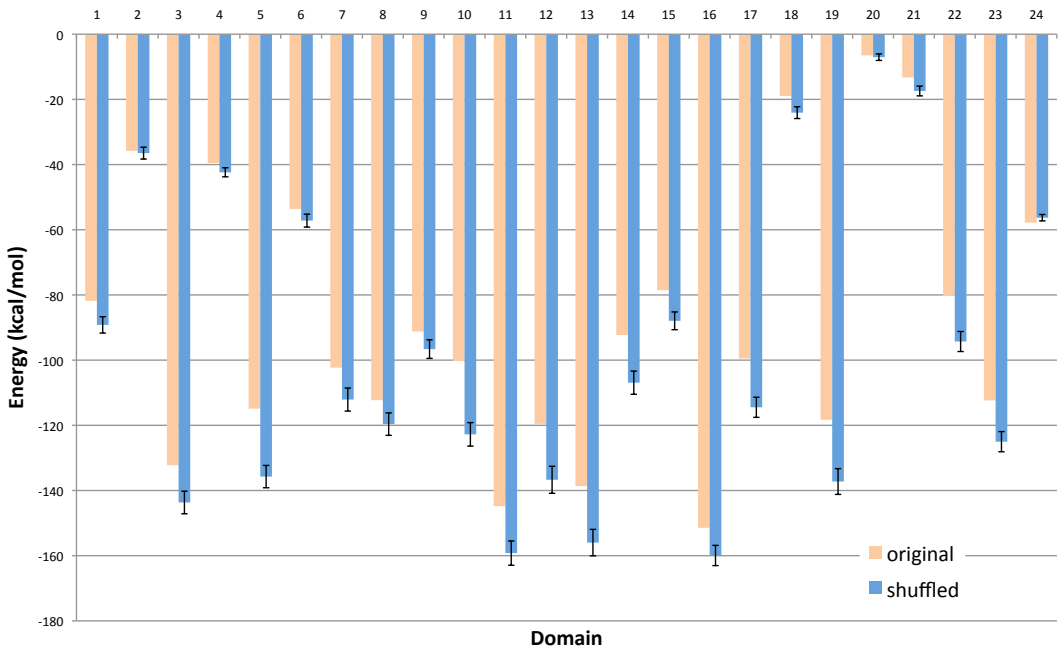

Figure 2.5: The energies of the MFE structures formed by the original (orange) and the shuffled (blue) sequences. The energy of the shuffled sequence for a given region is the average energy of 1000 structures formed by 1000 shuffled sequences.

Table 2.4: Z-score values calculated from the energies of the structures predicted from the original and same-composition shuffled sequences.

Region index	MFE	Centroid	Region index	MFE	Centroid
1	1.47	0.88	13	2.07	1.91
2	0.95	0.37	14	2.13	1.97
3	2.42	2.14	15	2.05	2.30
4	0.19	0.12	16	1.72	1.66
5	1.64	1.53	17	1.37	1.51
6	1.03	0.66	18	1.41	1.03
7	3.04	2.70	19	2.39	2.24
8	0.90	0.90	20	0.28	0.89
9	1.38	1.76	21	1.38	1.26
10	1.07	0.27	22	2.27	1.91
11	3.12	3.02	23	2.04	1.97
12	1.94	1.79	24	-0.76	-1.10
Averaged Z-score (MFE): 1.56					
Averaged Z-score (centroid): 1.40					

The enrichment of adenosines in the HIV-1 genome occurs largely in the wobble position of the codons

The apparent functional role of secondary structural elements (42) is perhaps the strongest evidence that the structure of Watts *et al.* is essentially correct. However, the

importance of the secondary structure raises an evolutionary problem for the virus. Any particular nucleotide faces pressures that favor particular protein sequences and, at the same time, pressures that favor particular secondary structures. How to reconcile these? Protein sequence is largely determined by the first two nucleotides in the codon, and there is more latitude in choosing the wobble nucleotide. This leads to the hypothesis that skewed compositions should be concentrated in the wobble position, if the A richness of HIV-1 RNA sequence is not just a protein coding requirement. To test the hypothesis, I examined the fractions of each nucleotide at the 1^{st}, 2^{nd} and 3^{rd} positions of the codons in all the open reading frames (ORF) of HIV-1 RNA. The results support the hypothesis: The third position of the codon has a higher proportion of adenosines (41%) than the first and second positions (34% and 33%, respectively) (Table 2.5), and for the single-stranded regions, the proportion of adenosines at the third position of the codon is even higher (54%). I also examined the codons where any change in the wobble nucleotide does not alter the coding (codons for Ser, Pro, Arg, Thr, Val, Ala, Gly and Leu). Again, the proportion of adenosines (48%) is significantly higher than that of any other nucleotide (G: 15%, C: 17%, U: 20%). In addition, to exclude the possibility that it is the tRNA abundance that causes the codon bias, I calculated the fractions of the nucleotides in the ORFs of Homo sapiens mRNA and compared the results to HIV. The data shows that the fraction of A at the 3^{rd} position of the codons for human is only 19%, meaning that the high concentration of A at the wobble position of the HIV-1 codons is not a result of human tRNA preference. Evidently, the virus exploits sequence flexibility in the wobble position to enrich nucleotides that favor the formation of particular secondary structures.

Table 2.5: The fractions of A, G, C, U at the three positions of the codons. The codons are collected from all the open reading frames from HIV-1 RNA.

	Fraction of nucleotide		
Nucleotide	1st position	2nd position	3rd position
A	0.34	0.33	0.41
G	0.30	0.21	0.21
C	0.19	0.20	0.15
U	0.17	0.26	0.23

SHAPE analysis does not appear to have missed significant regions of base pairing

The presence of long stretches of unpaired nucleotides in the secondary structure presented by Watts *et al.* is reminiscent of the early predictions for the secondary structure of the 16S ribosomal RNA, based on comparative sequence analysis (47). In the case of the rRNA, there were insufficient data to identify many of the base pairs, and "single-strandedness" was, at that time, a reflection of this fact, rather than a positive prediction that those nucleotides are not paired. I wondered if some of the unpaired regions in the structure of Watts *et al.* might actually contain base pairs that were missed by the original analysis.

To test this hypothesis, I folded all 24 of the domains identified in Table 2.3, using all base pairs in the published secondary structure as constraints. For each test, I made predictions using RNAfold and UNAFold. If there are any significant secondary structures that were missed in the published structure, they should appear as additional base pairs in the predictions from both programs.

Figure 2.6 shows a typical prediction for a particular domain from RNAfold. The program predicts only a handful of base pairs in addition to those in the structure of Watts *et al.* A similar number of additional base pairs are obtained when UNAFold is used to predict the structure, but most of those are different from those predicted by RNAfold. Figure 2.6 has three insets that compare the results of the two predictions. There is only one small double helix that appears in both predictions and where the SHAPE reactivities are reasonably low. There is no reason to believe that the SHAPE analysis has missed substantial regions of local secondary structure.

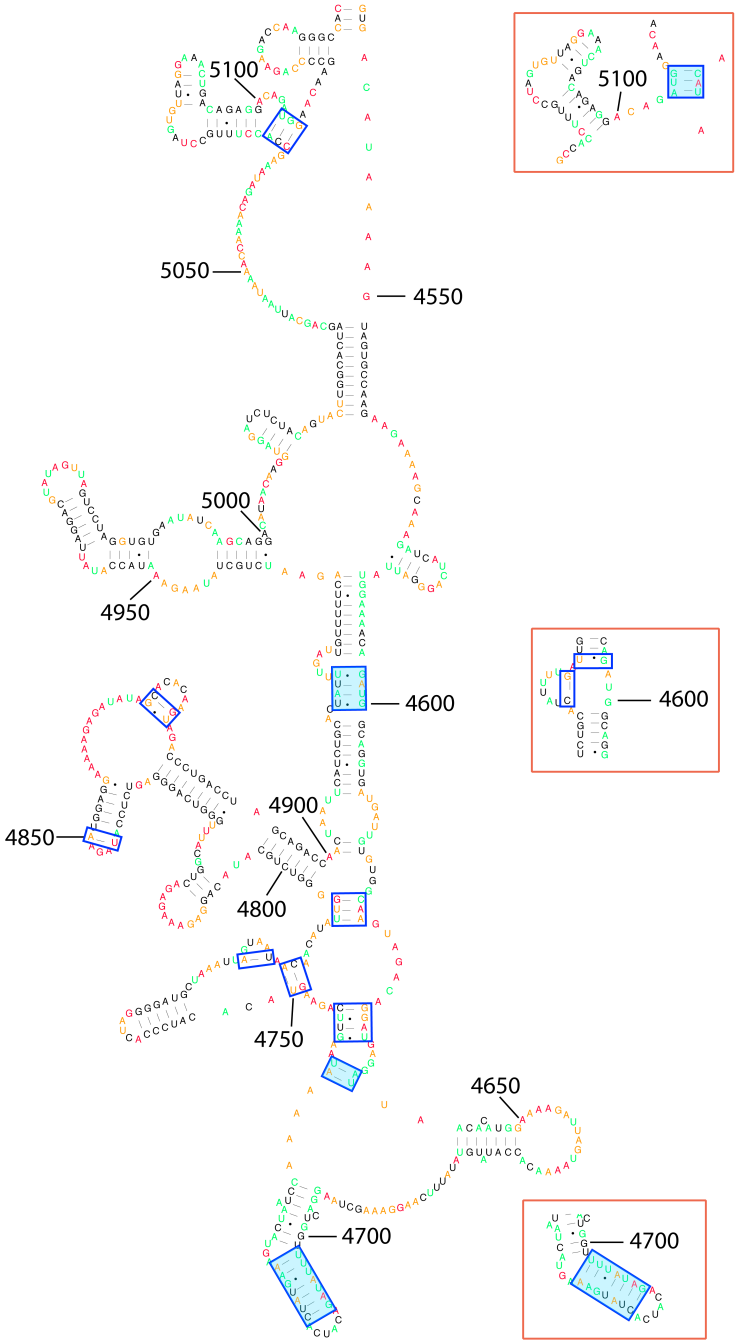

Figure 2.6: Secondary structure predictions for domain 11 (nt 3945-4518), using the base pairs in the published secondary structure (1) as constraints. The blue boxes identify base pairs that occur in the predicted structure, but not in the published structure; filled boxes indicate regions of lower SHAPE reactivities. UNAFold predicts only a small number of additional base pairs (main figure). The three insets show the predictions made by RNAfold for some regions, for sake of comparison.

In a similar fashion, I examined the collection of single-stranded regions that that run all the way from the 5'-end of the genome to the 3'-end and that connect successive domains. I wanted to see if there might be some long-range base pairing that was missed in the SHAPE analysis, since it was carried out with constraints that prohibited the formation of pairs between bases separated by more than 600 nucleotides. Replacing each of the 24 domains with a 12-nt stem-loop, I predicted the secondary structure for those inter-domain connectors. Both RNAfold and UNAFold predict some additional base pairs, and they agree on the prediction of five additional double helical regions (Figure 2.7). However, I do not regard these as strong predictions, because the SHAPE reactivities are rather high, and the double helices are rather weak, containing a substantial number of A-U and G-U base pairs. In addition, one stem joins regions that are over 5500 nucleotides apart in the primary sequence, which seems quite unlikely from an entropic point of view.

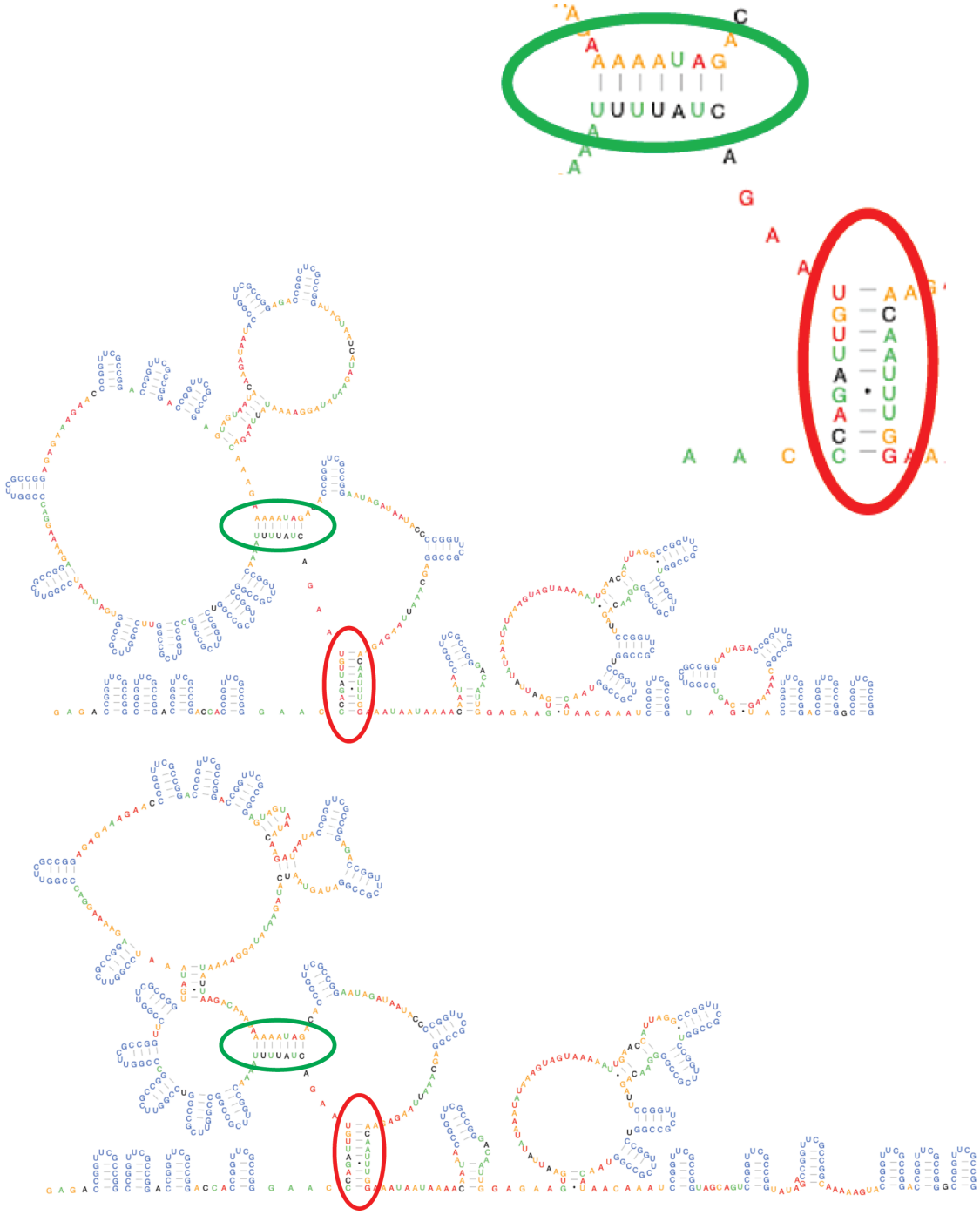

Figure 2.7: Secondary structures of inter-domain connectors predicted by UNAFold (above) and RNAfold (below). Blue stem-loops are replacements of the 24 domains. Red and green ellipses mark two examples of helices that appear in both predictions. All the predicted helices contain nucleotides with high SHAPE reactivities (colored in red and orange) and thus are considered as weak predictions.

In summary, the surprisingly large fraction of unpaired nucleotides is due to the highly skewed composition of the genome as a whole (A-rich), and of the single-stranded regions in particular. Local sequence effects also promote single-strandedness. The extremes of composition occur in the wobble position of the codon, allowing the virus to conserve important regions of single- and double-strandedness, while permitting protein sequences to evolve.

CHAPTER 3

A MODEL OF SATELLITE TOBACCO MOSAIC VIRUS

Abstract

Satellite tobacco mosaic virus (STMV) is an icosahedral T=1 single-stranded RNA virus with a genome containing 1058 nucleotides. X-ray crystallography revealed a structure containing 30 double-helical RNA segments, with each helix having nine base pairs and an unpaired nucleotide at the 3' end of each strand. Based on this structure, Larson and McPherson proposed a model of 30 hairpin-loop elements occupying the edges of the icosahedron and connected by single-stranded regions. More recently, Schroeder *et al.* have combined the results of chemical probing with a novel helix searching algorithm to propose a specific secondary structure for the STMV genome, compatible with the Larson-McPherson model. Here I report an all-atom model of STMV, using the complete protein and RNA sequences and the Schroeder RNA secondary structure. As far as I know, this is the first all-atom model for the complete structure of any virus.

Introduction

Satellite tobacco mosaic virus (STMV) is a T=1 icosahedral virus with a diameter of 17 nm (11,55,56). The genome of STMV is composed of a single-stranded RNA that has 1058 nucleotides. The 1.8 Å X-ray crystal structure of this virus identified 30 RNA double helices, one on each edge of the icosahedron (11) (Figure 3.1). Each helix is composed of 9 base pairs, plus an unpaired nucleotide at the 3' end of each strand. All

together the visible nucleotides account for 57% of the entire genome (11). The linkers between the helices and any RNA components in the interior of the virus are missing from the crystal structure, because the RNA does not have icosahedral symmetry; when the viral particle crystallizes, it can enter the lattice in any of 60 orientations, and the RNA density is averaged among these orientations. RNA sequence information is also absent in the crystal structure for the same reason.

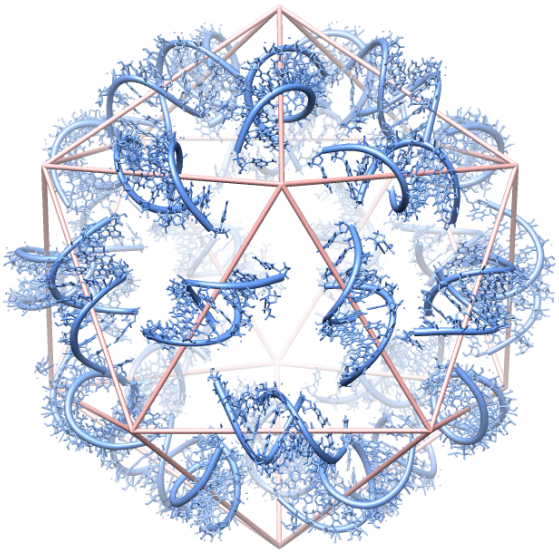

Figure 3.1: The crystal structure of STMV reveals thirty RNA double helices, each centered on a crystallographic two-fold axis.

Larson and McPherson proposed that the RNA is organized in a series of 30 hairpin loops that are linearly connected by single-stranded linkers (17). Recent atomic force microscopic images of STMV RNA shortly after release from the capsid are consistent with this linear arrangement of structural domains (57). Larson and McPherson presented a model in which the 30 hairpins situated on the 30 two-fold axes of the icosahedron are

connected by the shortest, most efficient path, suggesting both a structure for the mature virus and an efficient pathway for viral assembly; they then built a three-dimensional model of with 30 identical hairpin loops arranged along the proposed path (17). Since all RNA double helices appear identical in the icosahedrally averaged crystal structure, this model did not use the actual RNA sequence. Instead, each hairpin loop was composed of a stem of 9 A-U base pairs and a loop of 9 nucleotides. The stability of this model was subsequently demonstrated in molecular dynamics simulations (30).

There is also an all-atom model for a larger icosahedral RNA virus, Pariacoto Virus (T=3; 4322 nucleotides), developed by Devkota *et al.*(31). This model also used an artificial sequence designed to match secondary structure constraints derived from the previously published crystal structure (12).

Neither the previous model for STMV (30) nor the PaV model (31) contains the actual genomic sequence, so neither model examined the possible three-dimensional organization of secondary structures that might be formed by a naturally existing viral genome. It is now possible to do so, since Schroeder *et al.* have recently proposed a secondary structure for STMV RNA, based a combination of chemical probing and helix searching algorithms (58) (Figure 3.2). This secondary structure was designed to contain 30 local hairpin loops connected by single-stranded linkers, so it satisfies the Larson-McPherson motif.

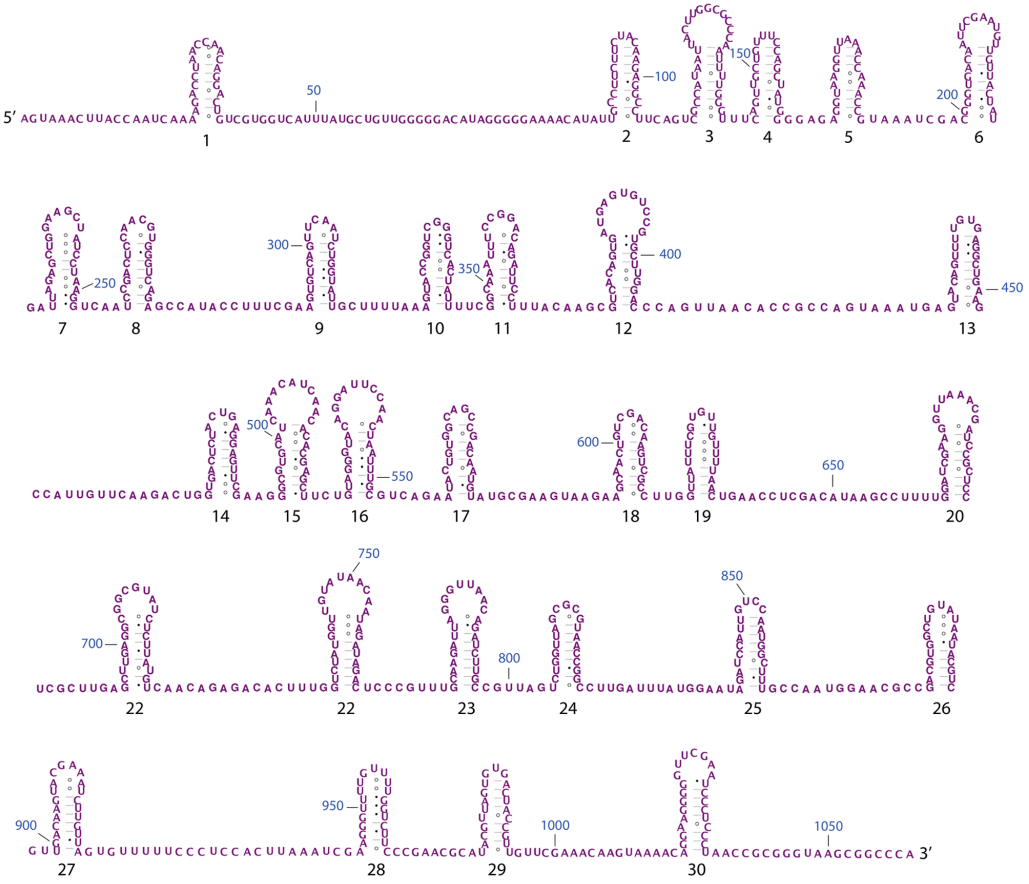

Figure 3.2: Secondary structure of STMV RNA determined by Schroeder *et al*. (58)

In this paper, I report an all-atom three-dimensional model of STMV. I demonstrate that STMV RNA adopting Schroeder's secondary structure (58) is able to cover all the edges of the icosahedron as originally proposed (17). Further, I compare the predicted electron density map for the model with maps derived from the crystal structure, finding that the model's RNA tertiary structure shares a strong similarity with that of the crystal structure.

Methods

The all-atom model for each hairpin loop was generated using MC-sym (59), based on its secondary structure. Each of these was minimized by NAMD (60) (CHARMM force field, conjugate gradient, 300 steps), and then superimposed onto the crystal structure (PDB ID 1A34) by matching the positions of corresponding phosphorus atoms, using VMD (61). The single-stranded connections were then built and linked to the hairpin loops using Sybyl-X (Tripos, St. Louis, Missouri). The one exception to this procedure was the very long connection between hairpin 1 and hairpin 2 (located in the viral interior), for which a coarse-grained three-dimensional model was generated using NAST (62), and then converted to all-atom structure by the program C2A (63).

After completion of the entire RNA model, the structure was minimized using NAMD, as described in the previous paragraph. The positions of P and C1' atoms in the double-helical stems of the crystal structure were used as restraints during minimization, assuring that the stems of the hairpin loops stayed aligned with those in the crystal structure. The restraints were set so that stronger restraining forces were assigned to positions with lower B-factors:

$$F = -2 \bullet [\ 8 - \mathrm{int}(B/30)\ ] \bullet x,$$

where x is the deviation of the atom from its ideal (crystallographic) position, measured in Ångstroms, and the force constant is measured in kcal/(mol•Å^2). The structure was then checked using the ADIT server (RCSB Protein Data Bank, http://deposit.rcsb.org/adit/). Chirality and penetration problems were fixed manually, and the structure was re-minimized. A few chirality and ring penetration

problems still remained, so manual manipulation and minimization were repeated a second time, giving a structure with no stereochemical violations. This structure was minimized to convergence (~13,000 steps.)

The coordinates of the protein subunit were obtained from the crystal structure (PDB ID 1A34). This subunit contains 147 residues, and it is missing the first 12 residues of the N-terminus. To add the missing residues, I used Sybyl-X (Tripos, St. Louis, Missouri) to build a tail of 12 residues following the actual peptide sequence in a way that the C-terminus of the tail was attached to the N-terminus of crystal structure. The whole capsid was generated from this 159-residue subunit using Oligomer Generator in VIPERdb (http://viperdb.scripps.edu/oligomer_multi.php). The protein capsid and RNA were then combined. After manually adjusted the positions of some tail residues to remove steric conflicts between the tails and the RNA, the complete structure was minimized with atoms restrained in a similar way as described earlier for the RNA alone (CHARMM force field, conjugate gradient, convergence in ~5000 steps). The restrained atoms included the P and C1' atoms of the RNA, and C, N and CA atoms of the protein residues obtained from the crystal structure (not including the tails). The tails were allowed free movement since their densities were missing in the crystal structure. ADIT analysis of the final model indicated that it is free of any serious steric conflicts or stereochemical violations, and that the chirality is correct for all chiral centers.

The molecular models were compared to the crystal structure by correlation of electron density maps (this comparison was done by Dr. Steven Larson at University of California, Irvine). The models were constructed with the crystal structure as the foundation, and therefore, they can be placed in a unit cell of P1 symmetry with cell

parameters **a**=174.27, **b**=191.77, **c**=202.50 Å and α=β=γ=90° to match the cell of the crystal structure. For the constructed models, occupancies and B factors for all atoms were assigned a value of 1.0 and 30 $Å^2$, respectively; for the models obtained from the crystal structure, the RNA strands have occupancies of 0.5 and B factors were reset to 30 $Å^2$. Structure factors, Fc, were calculated for each model. The structure factors were then used to calculate Fc electron density maps which were subsequently 60-fold averaged using the icosahedral symmetry operators to simulate the disorder that occurs by virtue of the fact that the virion can incorporate into the lattice in 60 different orientations. Dr. Larson chose to compare these model maps to two 2Fo-Fc maps based on a new refinement of the model against the original data deposited in the Protein Data Bank that is currently in progress (PDB ID 1A34) (64). These maps represent my best representation of the "true" electron distribution for the RNA. A 2Fo-Fc map was calculated from the structure factors obtained at the end of the refinement. A mask that covered the whole interior of the capsid was overlayed on the map and all points in the map that were outside the mask were set to zero. The resulting map represents only the density inside the capsid, which is mainly RNA density. This map was averaged in the same manner as the model maps. Correlation coefficients were calculated between these two maps as a measure of agreement between the "true" RNA electron distribution and the model derived maps (both non-averaged and averaged) (see Table 3.1).

Results and Discussion

Starting from the RNA secondary structure, I generated a tertiary structure for each hairpin loop, superimposed the stem of each of these onto one RNA double helix in the crystal structure, and connected the stems by single-stranded nucleotides. I also added the N-terminal amino acids missing from the capsid proteins in the crystal structure. Building the model involved a series of automated and manual methods. (See *Methods* for details.) The final model contains every single residue for both the RNA and protein components of the virus, and it compares very favorably with the structural model obtained by X-ray crystallographic methods.

RNA secondary structure

I used the secondary structure of STMV RNA published by Schroeder *et al.* (58), without including a possible tRNA-like structure at the 3'-end of the genome. Such structures have been found at the 3'-end of the genomes in several RNA viruses, including tobacco mosaic virus and turnip yellow mosaic virus (65). Aminoacylation experiments and sequence analysis suggest that the 3'-terminal 188 nucleotides of the STMV RNA also fold into a tRNA-like structure (66). Those authors argued that this part of the structure involves multiple pseudoknots and is terminated with an acceptor stem. However, the most energetically stable secondary structure with pseudoknots, predicted by pknotsRG (67), is completely different from this structure. Furthermore, this proposed set of pseudoknots is not consistent with either thermodynamic predictions or the experimental data (58): chemical probing showed 25 nucleotides to be single-stranded within this 188 nt region, but in the tRNA-like model, 7 of those 25 nucleotides are base-paired. This suggests that, if a tRNA-like structure exists in the 3'-region of the STMV RNA, it must

undergo re-folding during packaging. There is evidence that the RNA secondary structure inside a mature virus can be different from the unpackaged RNA (68).

I also generated a number of secondary structures based solely on thermodynamic considerations, to see if a structure with substantially lower free energy could be accommodated into the virus. A typical structure is shown in Figure 3.3. Although it has a much lower predicted free energy than the Schroeder structure (-302 kcal/mol *vs.* -160 kcal/mol), it presents serious problems for three-dimensional modeling. The stem-loops are much more varied in length than those in Figure 3.2, and there are many more bulged nucleotides and bent stems than those of the Schroeder structure. I built and refined an all-atom model using this alternative secondary structure, but it compared much less favorably with the crystallographic electron density than does the model using the Schroeder secondary structure.

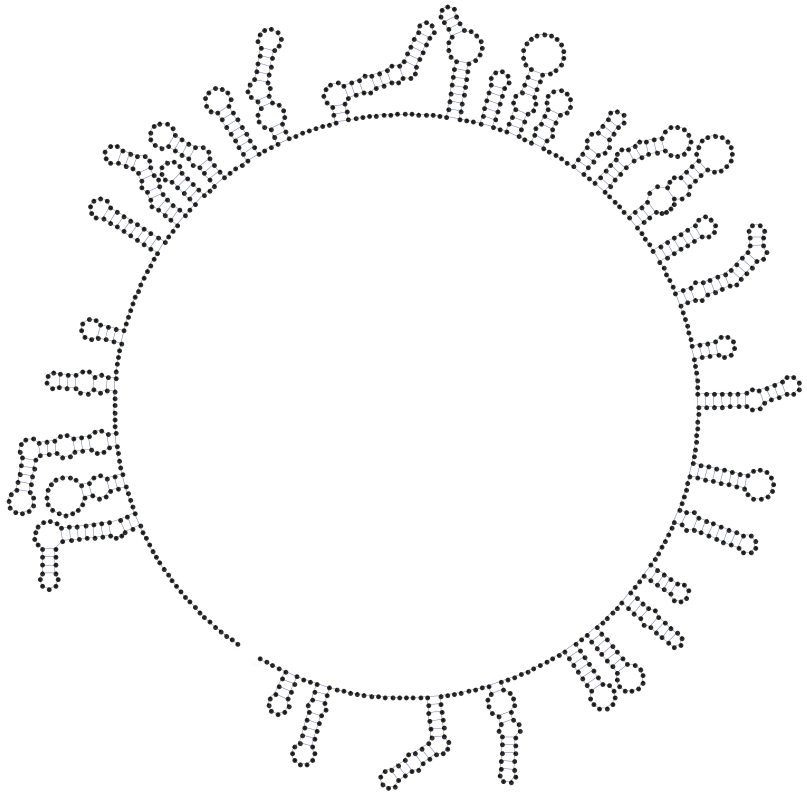

Figure 3.3: STMV RNA secondary structure predicted by UNAFold. Base-pairing distance is restrained to 45 nt.

Mapping the secondary structure onto the icosahedral geometry

The secondary structure in Figure 3.2 can be arranged with hairpin loops around the icosahedron so that each edge is covered by one hairpin loop (Figure 3.4). I arranged the hairpin loops on neighboring edges according to the lengths of the connections between them: for linkers of less than 8 nucleotides, the neighboring hairpins must adopt the tail-to-tail conformation, while for linkers longer than 13 nucleotides, head-to-tail or head-to-head conformations can be used as necessary. I located the long connection between hairpin loop 1 and hairpin loop 2 in the interior of the virus, rather than on the surface. This arrangement is necessary, because this connection is too large to be accommodated

right under the capsid. I emphasize that the arrangement shown in Figure 3.4 is just one of a very large number of possibilities.

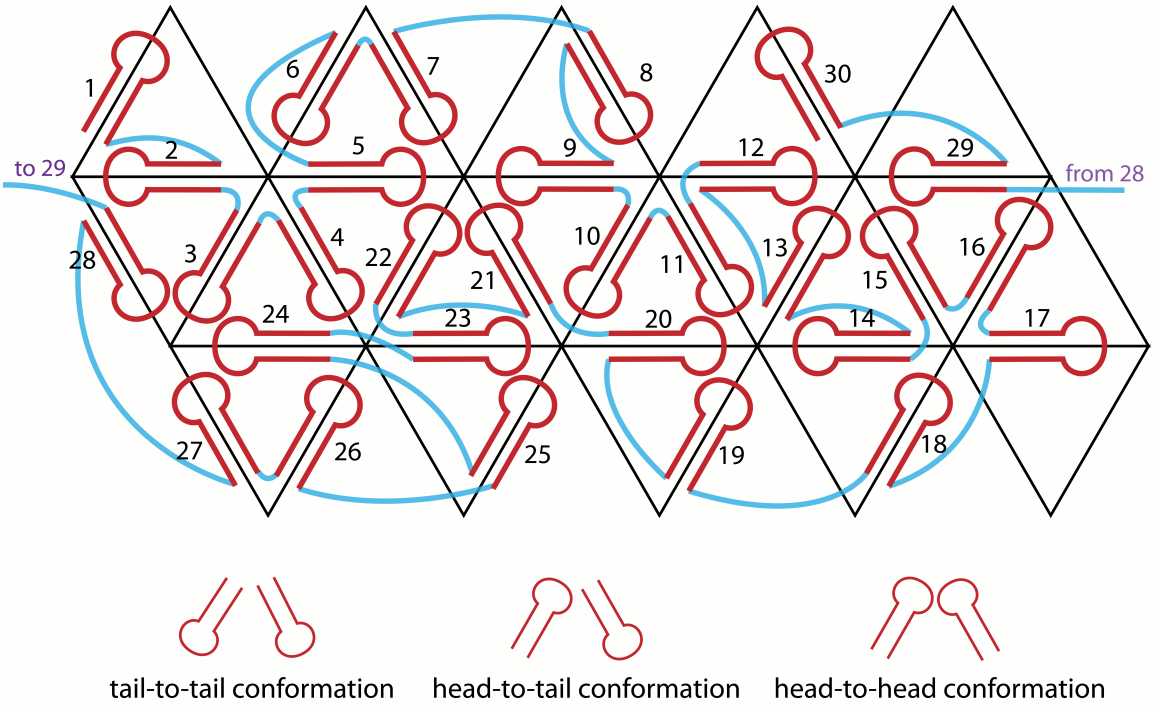

Figure 3.4: Implementation of the Larson-McPherson scheme (17) for mapping the hairpin loops (red) of the secondary structure in Figure 3.2 onto the icosahedron, with single-stranded connections colored in blue. Each triangle represents one face of the icosahedron.

The all-atom model

I built all-atom models for each of the hairpin loops in Figure 3.4 and superposed each of them onto one of the 30 helices in the crystal structure, then connected them with single-stranded nucleotides. This involved an iterative combination of manual and automated operations, including energy minimization (see *Methods* for details). The final model

contains all the protein subunits and the entire genome of 1058 nt (Figure 3.5). The RNA is free of any chirality problems and steric conflicts.

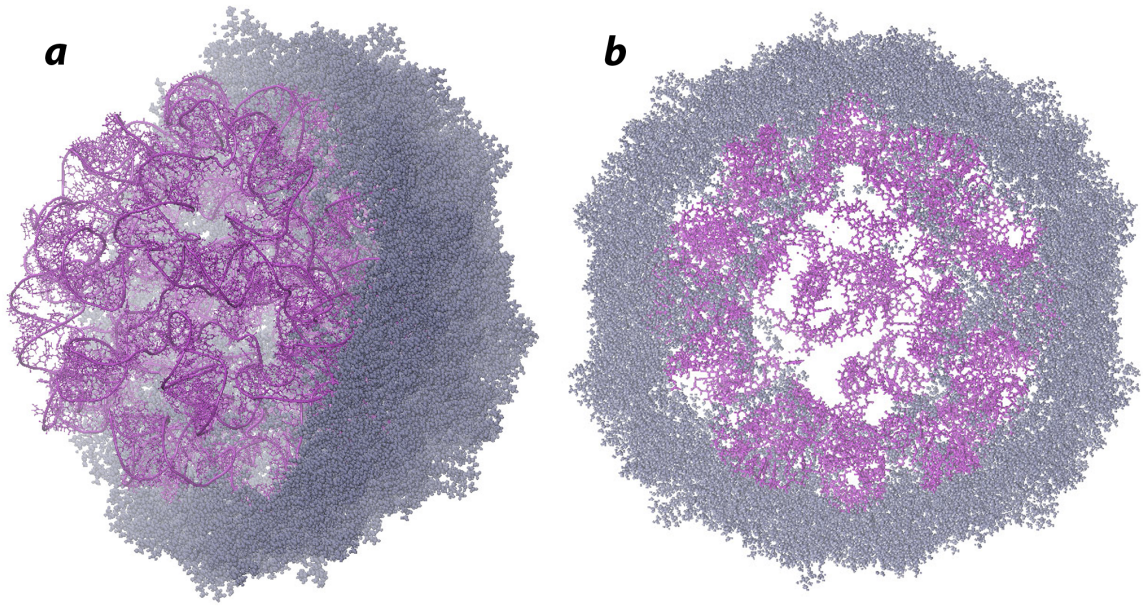

Figure 3.5: Final model of STMV (gray: protein; magenta: RNA) (a). Entire RNA with the protein capsid cut in half, to reveal the internal structure. (b). A 50 Å section through the center of the final model.

To test the quality of the model, Dr. Steven Larson, our collaborator, compared it with the original X-ray crystallographic data, and he did the same with the model previously published by Larson and McPherson (17) and examined in MD simulations by Freddolino *et al*.(30). To begin with, the RMSD between the phosphorus atoms of the model and those of the crystal structure is only 1.21 Å (Figure 3.6). An electron density map for each model was created and compared to the 2Fo-Fc map derived from the crystal structure. The maps of my model are much better correlated with the 2Fo-Fc map

than the older model, with significantly higher correlation coefficients for both the full RNA model and for the helical regions alone (Table 3.1). The correlation coefficient for my model (0.557 with icosahedral averaging) is 91% of 0.611, the correlation coefficient for the crystallographic model (PDB ID 1A34) The excellent agreement is clearly seen when the model is superposed on the 2Fo-Fc electron density map (Figure 3.7).

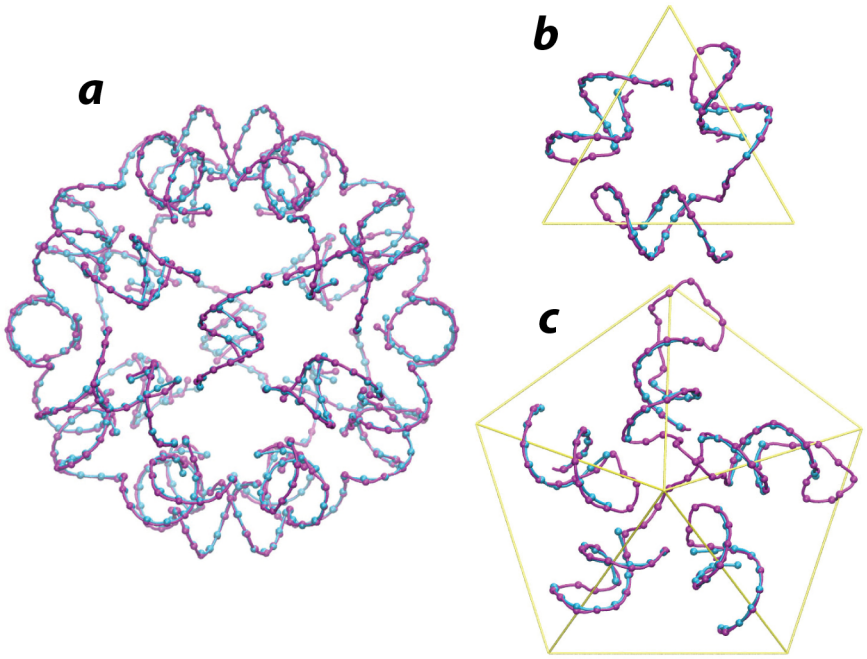

Figure 3.6: Hairpin loops in the model shown together with the helices in the crystal structure (magenta: model; blue: crystal structure). A. All 30 hairpins (loops are not shown for visual clarity). B. View along one of the 5-fold axes. C. view along one of the 3-fold axes.

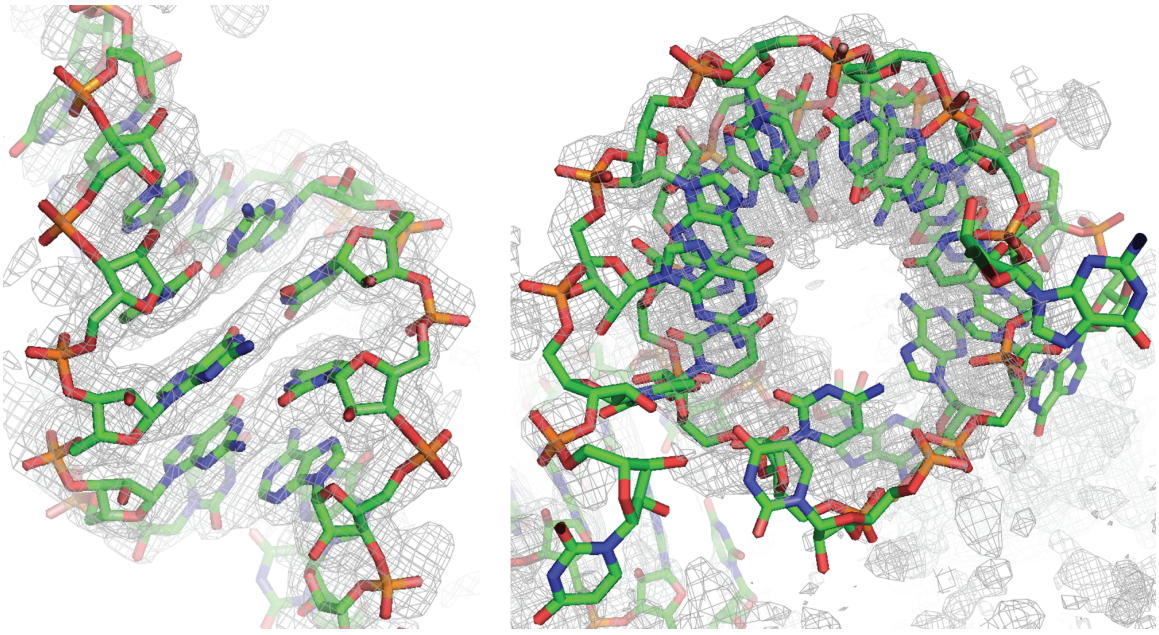

Figure 3.7: Two views of one helix from the current model, superimposed on the crystallographic 2Fo-Fc map. The model has thirty unique double helices, and the figure shows a typical level of agreement between the model and the experimental data.

Table 3.1: Correlation coefficients between the crystallographic electron density and the predicted electron densities from the RNA models (2Fo–Fc maps), with areas where the correlation coefficient exceeds 0.5 highlighted in grey.

	X-ray RNA model		Previously published model (Larson and McPherson, 2001)		Current RNA model	
	unaveraged	averaged	unaveraged	averaged	unaveraged	averaged
Full model			0.231	0.391	0.299	0.524
Helices	0.521	0.611	0.288	0.443	0.388	0.557

Twelve residues of the positively charged tails at the amino terminus of the capsid proteins are not visible in the crystal structure, presumably because the 60 tails have a variety of conformations; they are statically disordered. There is not sufficient experimental information to allow us to predict those conformations with any accuracy, but the tails in my model do penetrate through the array of RNA double helices just under the capsid, reaching toward the center of the virus (Figure 3.8). In building my earlier model of PaV (31), I forcibly stretched the tails toward the center of the virus. I made no such effort in the STMV model reported here, nor did I attempt to position the tails so their positive charges would maximize neutralization of the RNA charge. As a consequence, the results shown in Figure 3.8 represent a first-order approximation to the tails' positions, not a set of specific predictions.

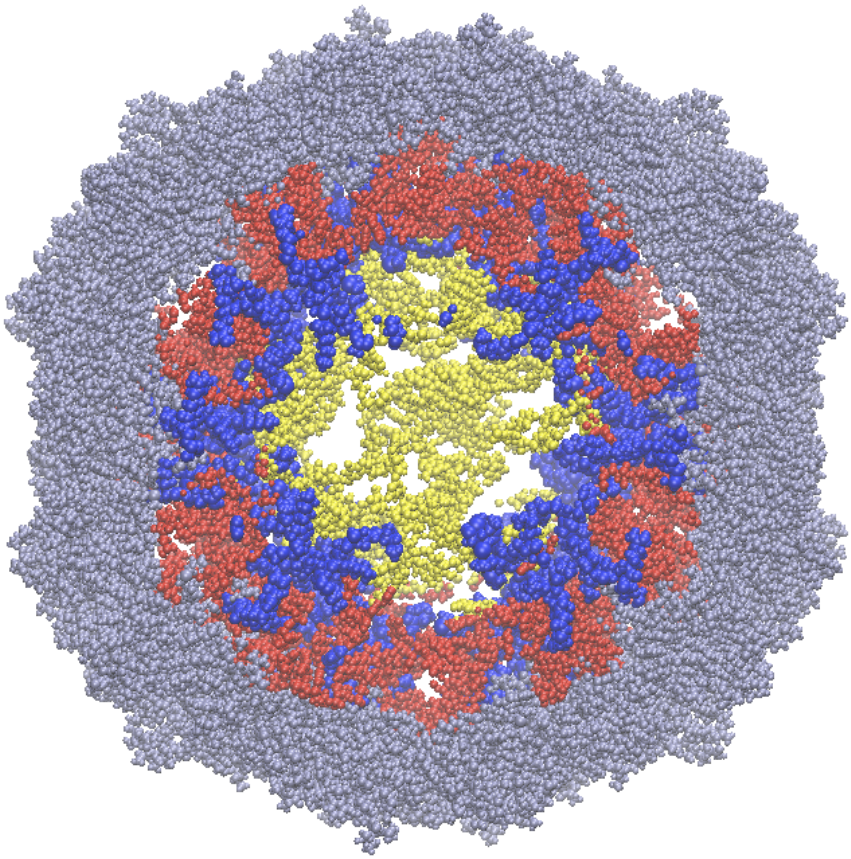

Figure 3.8: The 12 residues of the N-terminal protein tails (blue) pass through those double-helical regions of the model RNA (red) that correspond to the double helices seen in the crystal structure. The figure shows a slice (section) through the center of the virus; the RNA section is 50 Å thick, while the protein section is 60 Å thick, to facilitate visualization of the tails. The crystallographically observed regions of the capsid proteins are grey, and those parts of the RNA model that do not correspond to regions in the crystal structure are yellow. For scale, the virus is about 165 Å in diameter. I emphasize that there is no experimental information on how deeply the protein tails reach into the center of the virus, and that I have not attempted to stretch the tails. Fully extended, 12 amino acids could just reach the center of the virus.

As far as I am aware, this is the first all-atom model for any virus that is based on the actual sequence of the genome. The model demonstrates that the secondary structure proposed by Schroeder *et al.* (58) can be realized in three dimensions, providing support both for that structure and for the original suggestion that the genomic RNA is organized in a series of 30 stem-loops (17). The very high correlation between the model and crystallographic electron density maps argues that I have captured the essential features of the actual structure. An alternative secondary structure with a lower folding free energy has much less regular stem-loop structures than does the Schroeder model, and it has a much lower correlation with the experimental electron density. This argues that the secondary structure of the packaged RNA may be different from that of the free RNA, and the capsid could play a substantial role in directing RNA refolding.

This model raises two major challenges for future research. First is the question of whether viral assembly would be favored or hindered by structural heterogeneity. It is not known if all STMV genomes in the mature virus have essentially identical secondary structures. Even if they do, the three-dimensional organization shown in Figure 3.4 is not unique, and this might vary from virion to virion. The second challenge is to understand the structural, thermodynamic and kinetic aspects of viral assembly. Theses issues merit experimental attack, and I am developing models for simulating the assembly process and examining these questions computationally.

CHAPTER 4

A MODEL OF BACTERIOPHAGE MS2 AND ITS IMPLICATIONS IN VIRAL GENOME PACKAGING

Abstract

Bacteriophage MS2 is a T=3 icosahedral virus, whose genome contains a single piece of RNA with 3569 nucleotides. The cryo-EM density obtained by Toropova and coworkers revealed a double-shell structure of the RNA. To explore the three-dimensional organization of the genome in detail, I studied the possible secondary structures of the RNA and built an all-atom model of the virus. The final model successfully captures the double-shell feature presented in the cryo-EM density, and the similarity that I found between the RNA secondary structures predicted *in vitro* and *in vivo* suggests that the genome packaging occurs after the entire RNA has been synthesized. I also present evidence that the RNAs of viruses requiring packaging signals have evolved to be structurally compact, which facilitates post-replicational RNA packaging. In contrast, viruses that do not depend on packaging signals probably adopt co-replicational RNA packaging.

Introduction

Bacteriophage MS2 (MS2) is a T=3 icosahedral RNA virus. The genome is composed of a positive sense RNA with 3569 nucleotides (15). Each viral particle has 180 copies of the protein subunit, which adopt three conformations, designated as the A, B and C forms

(15). In the absence of the RNA, every two protein subunits gather into a C-C dimer, and upon binding to an RNA stem-loop, the C-C dimer is transformed into an A-B dimer (allosteric switching) (13). The capsid of the mature virus contains 60 A-B dimers and 30 C-C dimers (Figure 4.1). In addition to the capsid protein, the virus also has one copy of maturation protein packaged inside the capsid (15,69). MS2 infects male *E.coli* by attaching to the F-pilus of the bacteria and injecting its RNA together with the maturation protein into the cell (15).

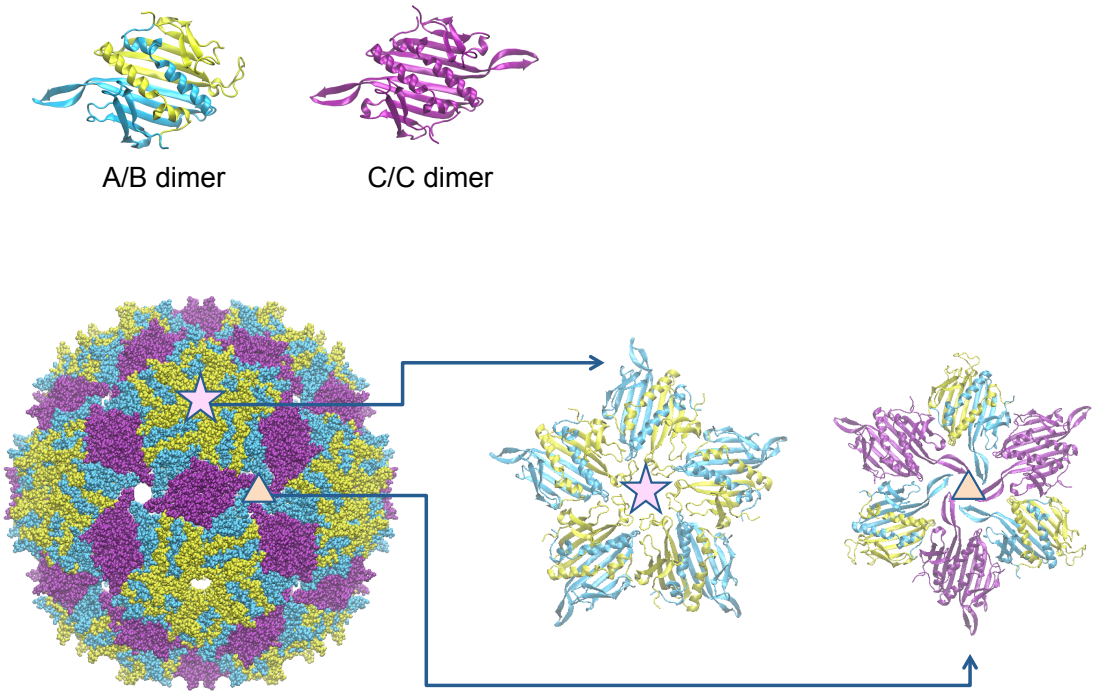

Figure 4.1: Capsid protein of MS2 (star: 5-fold axes; triangle: 3-fold axes)

The assembly of MS2 requires a packaging signal, which is an RNA stem-loop formed by a 19-nt sequence (TR) (Figure 4.2) located in the middle of the entire genome (1746-

1764) (16). The binding of a C-C dimer to the TR sequence represses translation and initiates the self-assembly process. This type of sequence-specific packaging signal (PS) exists in many other viruses as well, including satellite tobacco necrosis virus (STNV), brome mosaic virus (70), hepatitis B virus (71) and influenza virus (3). On the other hand, there are viruses that apparently do not require such signals; and among these are satellite tobacco mosaic virus (STMV) (17) and cowpea chlorotic mottle virus (CCMV) (72). It remains unknown whether PS-positive viruses have different packaging pathways from PS-negative ones. Two groups working on STMV, a PS-negative virus, have predicted secondary structures for the genomic RNA (44,58), and a significant difference was observed between the *in vitro* (44) and *in vivo* (58) structures. There are also computational methods to analyze viral RNA secondary structures. For example, Yoffe *et al*. introduced a metric for measuring the extendedness of RNA secondary structure, the maximum ladder distance (MLD) (43). MLD is the maximum value of ladder distance, *LDij*, for all combinations of *i* and *j*, where *LDij* is the number of base pairs that are crossed along the most direct path from base *i* to base *j* in the two-dimensional secondary structure graph. Here I examine the structure of MS2 RNA to see how it might facilitate packaging and how it might differ from PS-negative viral RNAs.

Various experiments, including chemical probing, X-ray crystallography and cryo-EM, have led to a better understanding of the structures of MS2 RNA. The secondary structure of the RNA in the absence of the capsid was determined using chemical probing during the 1980s (unpublished). In terms of the tertiary structure, X-ray crystallography studies of viral-like particles composed of viral protein and artificial RNA proved that RNA stem-loop variants are able to bind to the protein dimers (16). The crystal structures of

those viral-like particles, together with the fact that RNA stem-loop triggers the allosteric switching of the protein dimer from C-C to A-B form, suggest that the RNA of the wild-type virus forms a series of stem-loops that bind to the capsid protein. The locations of those stem-loops in the genome were predicted by Stockley *et al.* using SELEX (Figure 4.2). In addition, an icosahedrally averaged cryo-EM density at 9 Å resolution revealed that the wild-type MS2 virion has its RNA organized into a double-shell structure, with densities underneath both A-B and C-C dimers, as well as along the 5-fold axes (14). Another cryo-EM image obtained from 5-fold rotational averaging at 20 Å resolution revealed the asymmetric distribution of the RNA and the location of the maturation protein (29). Based on the cryo-EM density, Dykeman et al. proposed a three-dimensional layout of the RNA, which follows a Hamiltonian path and is consistent with the electron density (73). Collectively, these data offer the opportunity to model the wild-type MS2 RNA structure at atomic level, hopefully providing insights into how the genome folds into such a complicated conformation.

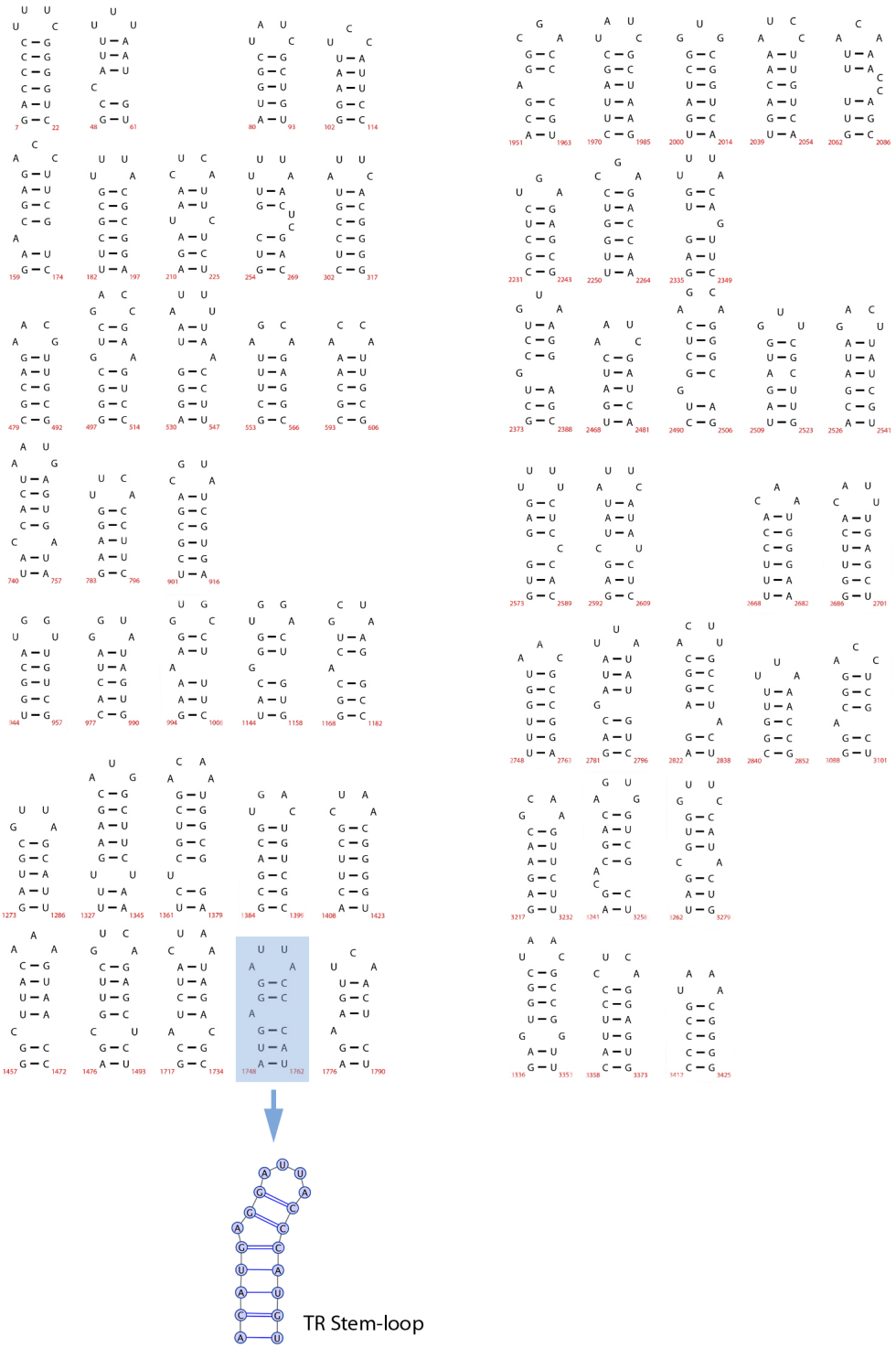

Figure 4.2: 60 stem-loops predicted using SELEX. (Stockley *et al*., unpublished) The highlighted stem-loop is part of the TR stem-loop.

In order to explore the organization of the genome in detail, I built an all-atom model of MS2. Previously, I had constructed an all-atom model of STMV, a T=1 RNA virus, using the real sequence of the genome (74). This model was based on the assumption that the RNA is likely to form a high-energy secondary structure of a series of local stem-loops during assembly, which further suggested that protein-RNA interactions are the major forces that shape the RNA secondary structure *in vivo* (58). To examine whether the formation of MS2 RNA structure depends largely on the capsid protein, I started from 60 stem-loops located using SELEX (unpublished) and predicted a secondary structure for the entire MS2 RNA. As will be seen, I conclude that MS2 uses a different packaging strategy than STMV.

Methods

RNA secondary structure prediction and analysis

Starting with the 60 stem-loops shown in Figure 4.2, I predicted the secondary structure of each domain that connects two consecutive stem-loops using UNAFold (34). For each domain, I selected one of the three structures that were predicted to have the lowest free energies (most of the structures selected are MFE structures except that two domains have the structures with the second lowest free energy and one domain with the third lowest free energy). Those domains were then combined with the 60 stem-loops to form the secondary structure for the entire RNA.

To examine the compactness of the RNA, I calculated the maximum ladder distance (MLD) (43) of the secondary structure that I predicted, as well as the MLD of the structure determined using chemical probing. In addition, I generated 100 shuffled RNA

sequences with the same composition as the MS2 genome, then predicted the MFE structure of each of the shuffled sequence using RNAfold and calculated the MLD for each of them. To compare the results with other viruses, I did the same MLD analysis for the genomes of several other viruses, including both PS-positive (PaV, STNV) and PS-negative ones (STMV). For the viruses whose RNA secondary structures have not been probed, I predicted the MFE structures using RNAfold.

Construction of the tertiary structure of the RNA

I generated all-atom models for each of the 60 stem-loops in my secondary structure model, using MC-sym (59). Then, I superimposed each of those onto the RNA stem-loops from the crystal structure of the viral-like particle composed of the MS2 protein capsid and 60 copies of the TR stem-loop (PDB ID: 1ZDK) (16) (Figure 4.3). Before the superimposition, I used the MDFF plugin in VMD (61) to adjust the positions of the stem-loops in the crystal structure by fitting them into the cryo-EM density obtained from icosahedral averaging (PDBe ID: EMD-1431) (14).

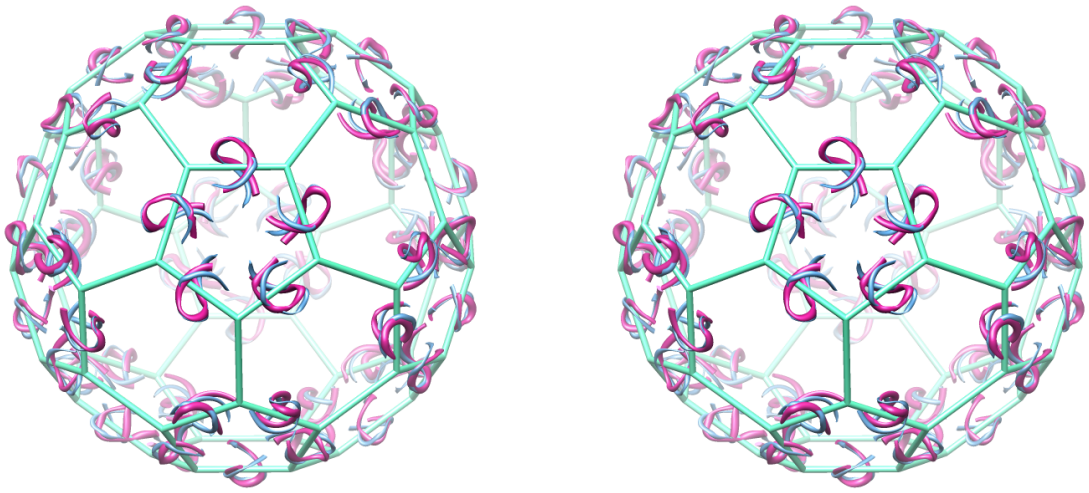

Figure 4.3: 60 stem-loops superimposed onto the crystal structure (1ZDK) (purple: model; blue: crystal structure)

For the domains connecting the stem-loops, I converted the secondary structures into all-atom models, using MC-sym and RNAComposer (75). For single-stranded connections, I generated all-atom models using Sybyl-x (Tripos, St. Louis, Missouri). I based the locations of the domains inside the virus on the Hamiltonian path proposed by Dykeman *et al*. (73) (Figure 4.4), and each domain was placed either along the 5-fold axis or underneath the C-C dimers, according to their assigned locations in Figure 4.4. After all the domains were in place, I connected them with the 60 stem-loops using VMD, forming the complete sequence of 3569 nucleotides with no breaks. I then fitted the entire RNA into the cryo-EM density using the MDFF plugin in VMD (Figure 4.5), with a harmonic restraint (force constant: 200kcal/(mol•Å^2)) placed on each of the atoms of the previously placed 60 stem-loops to restrict their movement.

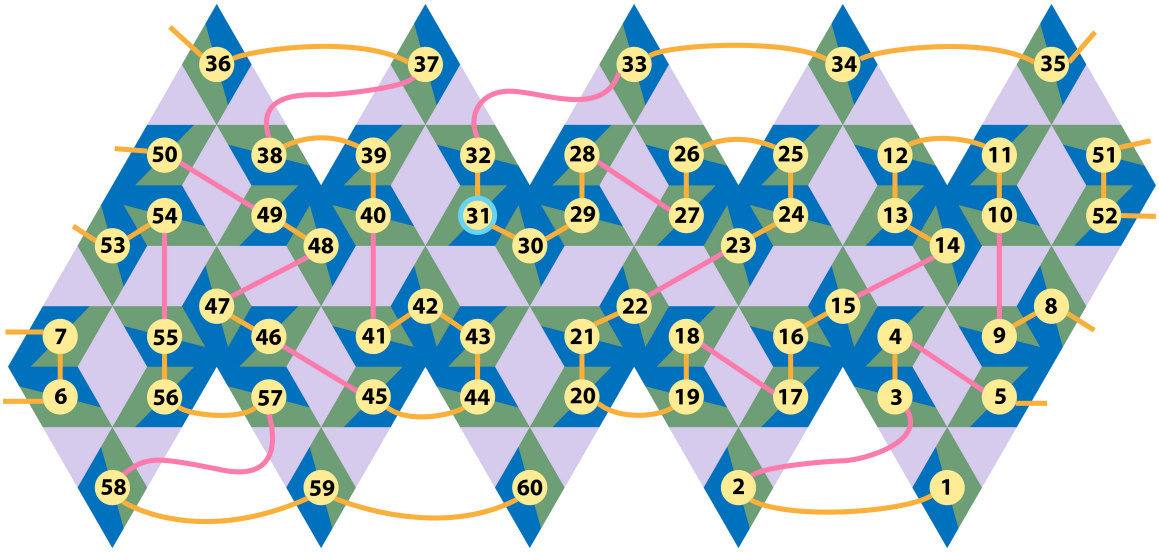

Figure 4.4: Layout of MS2 RNA as a Hamiltonian Path (orange lines: short connections; pink lines: long connections underneath C-C dimers; circles: the 60 stem-loops; circle 31: TR stem-loop)

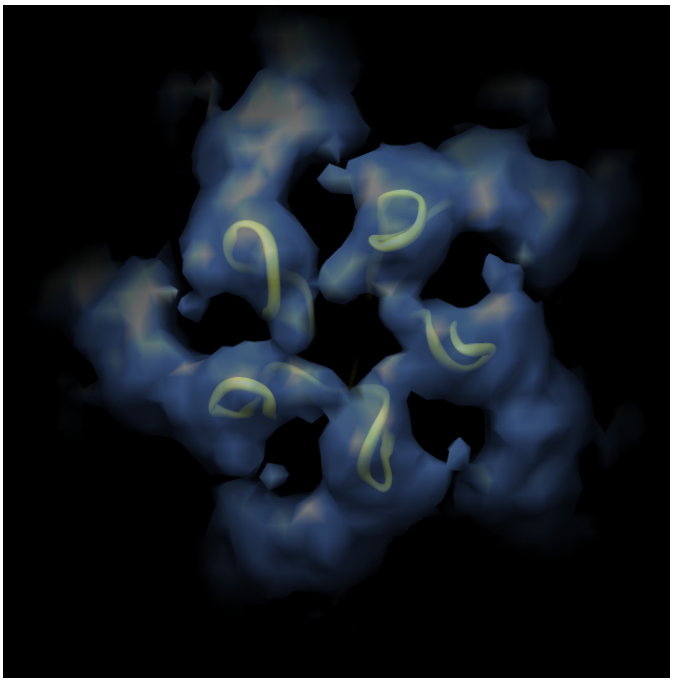

Figure 4.5: MS2 RNA fitted into the cryo-EM density (view along 5-fold axis)

Construction and docking of maturation protein

I predicted the tertiary structure of the maturation protein using TASSER (76), and placed it near the two ends of the RNA, a location that was indicated in the cryo-EM density obtained from 5-fold rotational averaging. The whole model, including the RNA and the maturation protein, was minimized using NAMD (conjugated gradient, 5000 steps) (60).

Results and Discussion

Secondary structures of MS2 RNA

The secondary structure that I predicted for MS2 RNA (Figure 4.6) shares strong similarities with the structure of the same RNA predicted using chemical probing in its unpackaged state. The comparison between the two structures shows that all the stem-loops in the structure probed *in vitro* are re-captured in my prediction, and the structures of the domains that connect the SELEX-based stem-loops are highly consistent with those of the corresponding regions in the chemically-probed structure (Figure 4.7). This structural resemblance suggests that the secondary structure of the RNA in the mature virus is largely formed before it is packaged into the capsid, and this secondary structure is not significantly changed upon protein binding. This result contrasts with the significant difference discovered between *in vitro* and *in vivo* STMV RNA structures (44,58). In the case of STMV, the RNA in the absence of the protein capsid has a very extended secondary structure, while the predicted RNA structure in the mature virus, which is supported by the all-atom model (74), is composed of a number of local stem-

loops and is free of any long-range base pairs. The disparity between MS2 and STMV suggests that they adopt different packaging strategies, which will be discussed later.

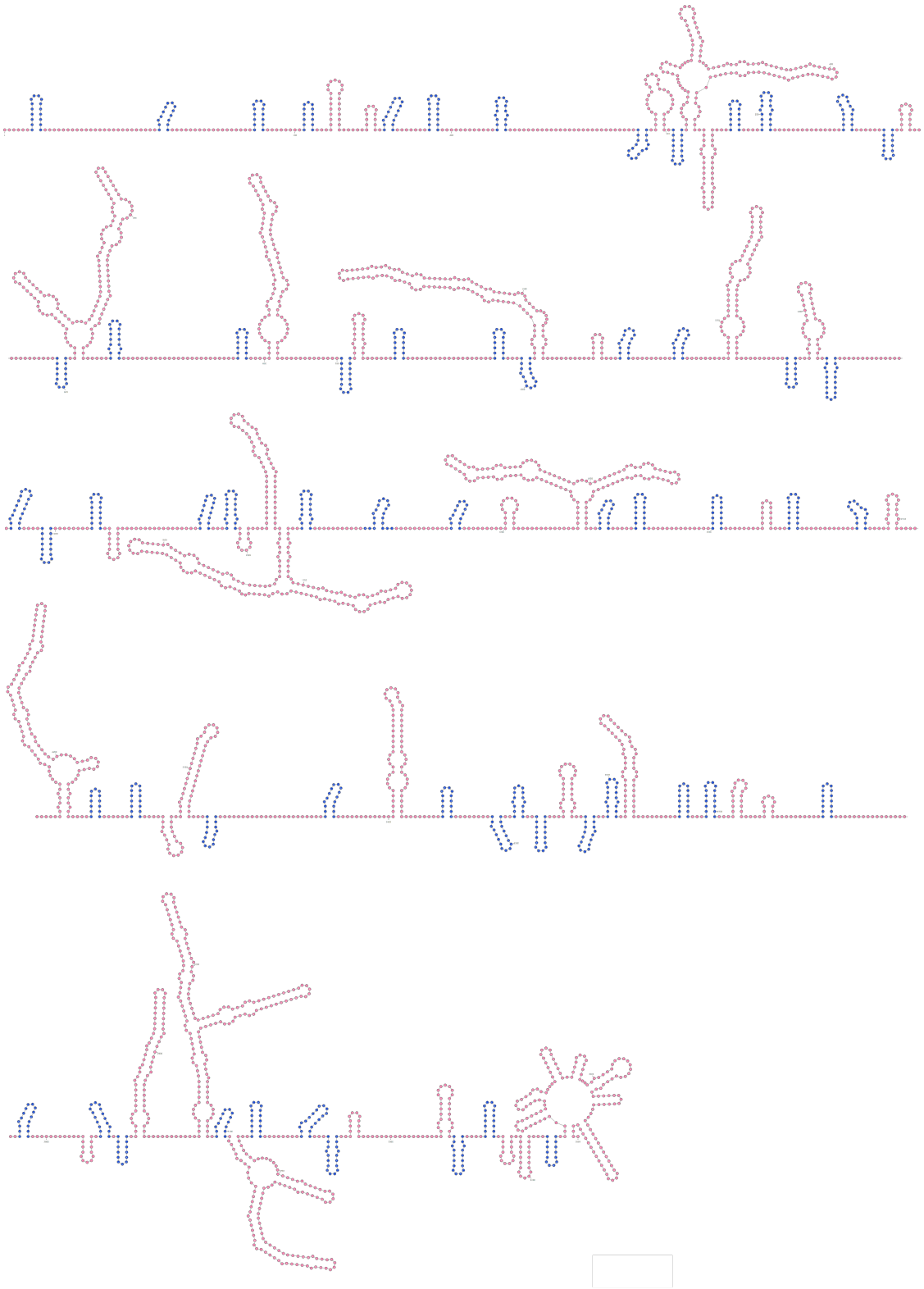

Figure 4.6: Predicted MS2 RNA secondary structure (blue: stem-loops located by SELEX; red: domains folded using UNAFold)

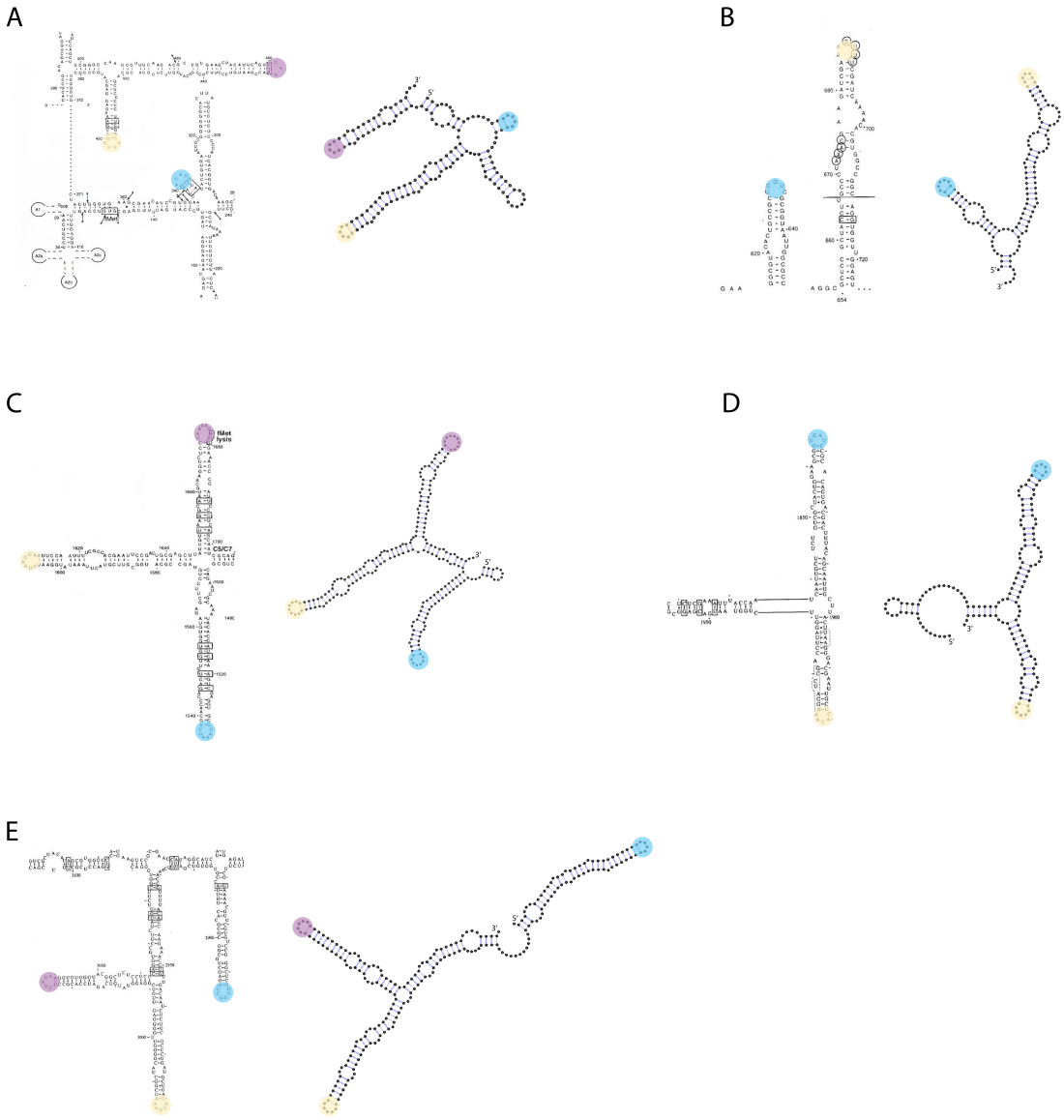

Figure 4.7: Comparisons between probed and predicted structures. (left: probed structure; right: predicted structure) A. Domain between stem-loops 9 and 10 (318-478) B. Domain between stem-loops 14 and 15 (607-739) C. Domain between stem-loops 29 and 30 (1494-1716) D. Domain between stem-loops 32 and 33 (1791-1950) E. Domain between stem-loops 53 and 54 (2853-3087)

Three-dimensional organization of the genome

The final model contains the entire genome and the maturation protein, which is located near the 3' and 5'-ends of the RNA and along the 5-fold axes (Figure 4.8). The model of the RNA has the double-shell features as shown in the cryo-EM density, and the locations of the shells are consistent with the experimental data (Figure 4.9). The radial density distribution calculated from the model clearly shows two peaks, corresponding to the inner shell (42-65 Å) and outer shell (84-108 Å) in the cryo-EM density.

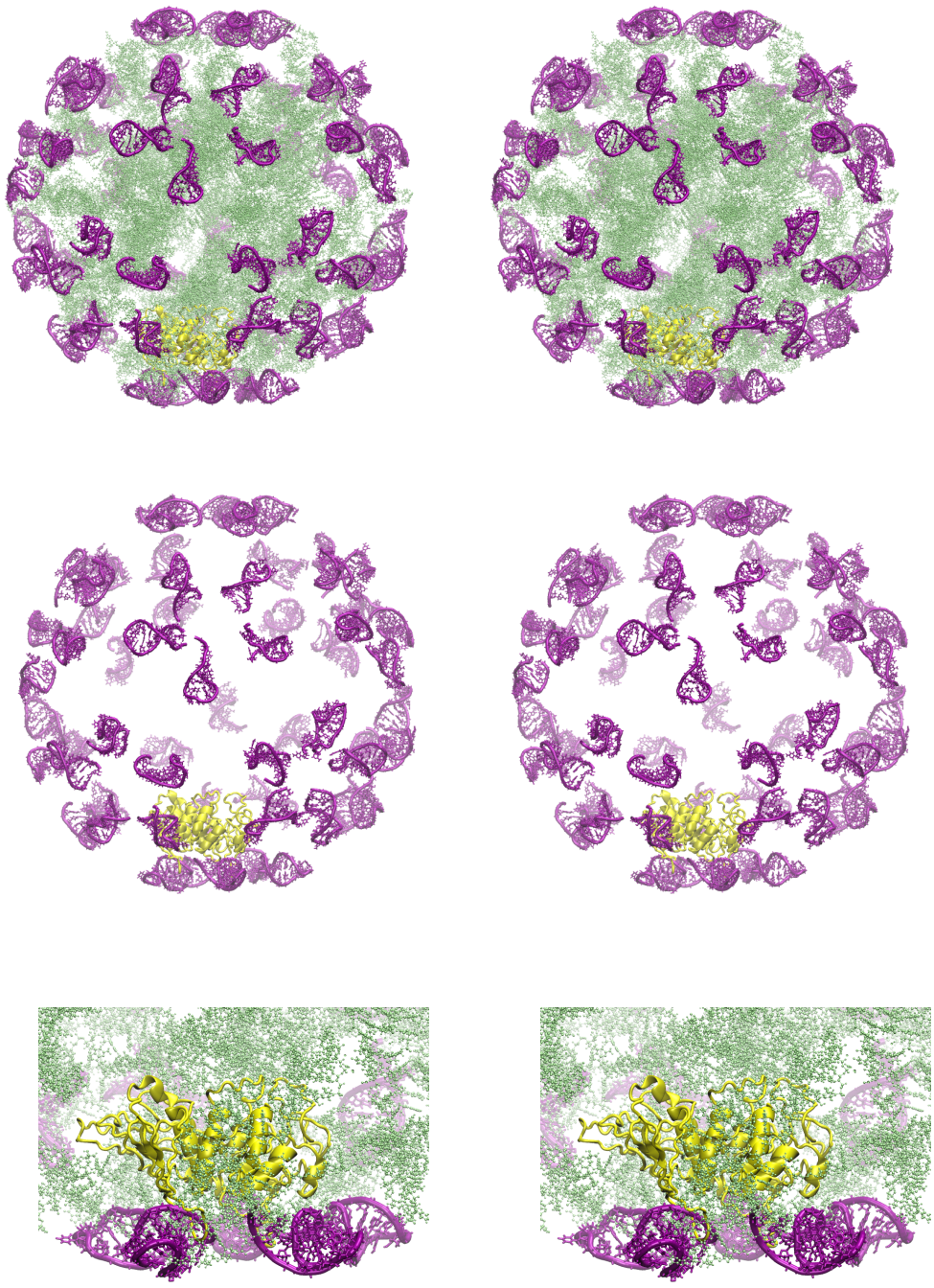

Figure 4.8: MS2 RNA and maturation protein (purple: 60 stem-loops; lime: the rest of the RNA; yellow: maturation protein)

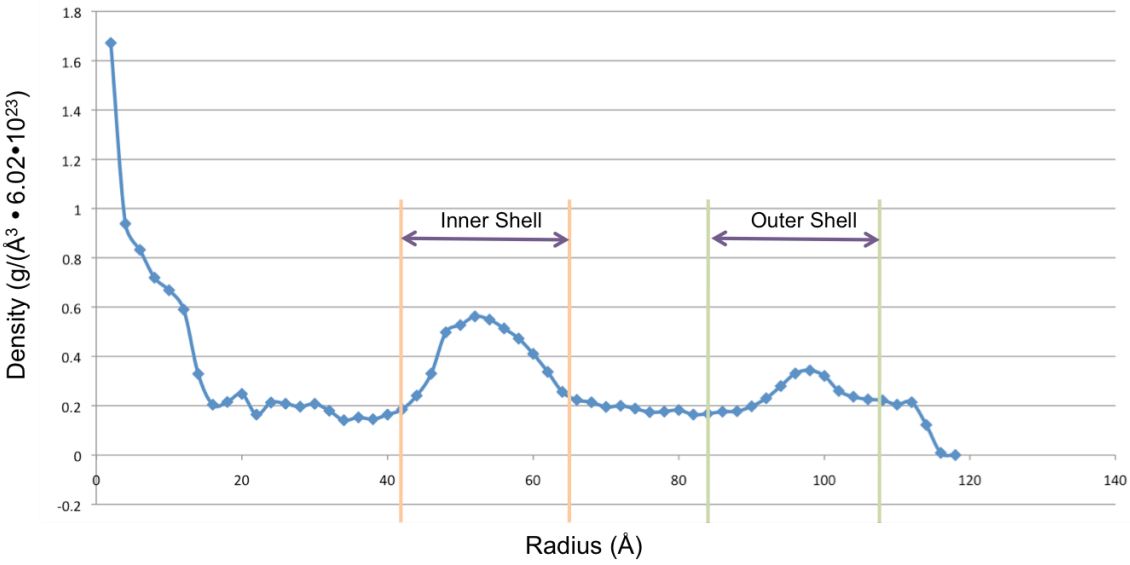

Figure 4.9: Radius density distribution of MS2 RNA in the model (inner and outer shell ranges are from the cryo-EM data)

The layout of the genome follows the predicted Hamiltonian path, and the success in capturing the double-shell feature of the cryo-EM density supports the validity of this path, as well as the secondary structure predicted using SELEX and UNAFold. In addition, I constructed three-dimensional models for the MS2 and STMV RNAs probed *in vitro*, and the comparisons with their *in vivo* models demonstrate that the difference in compactness between *in vivo* and *in vitro* MS2 RNA is much smaller than that between *in vivo* and *in vitro* STMV RNA (Figure 4.10), which is consistent with my findings from the secondary structure analysis (Figure 4.6 and Figure 4.7).

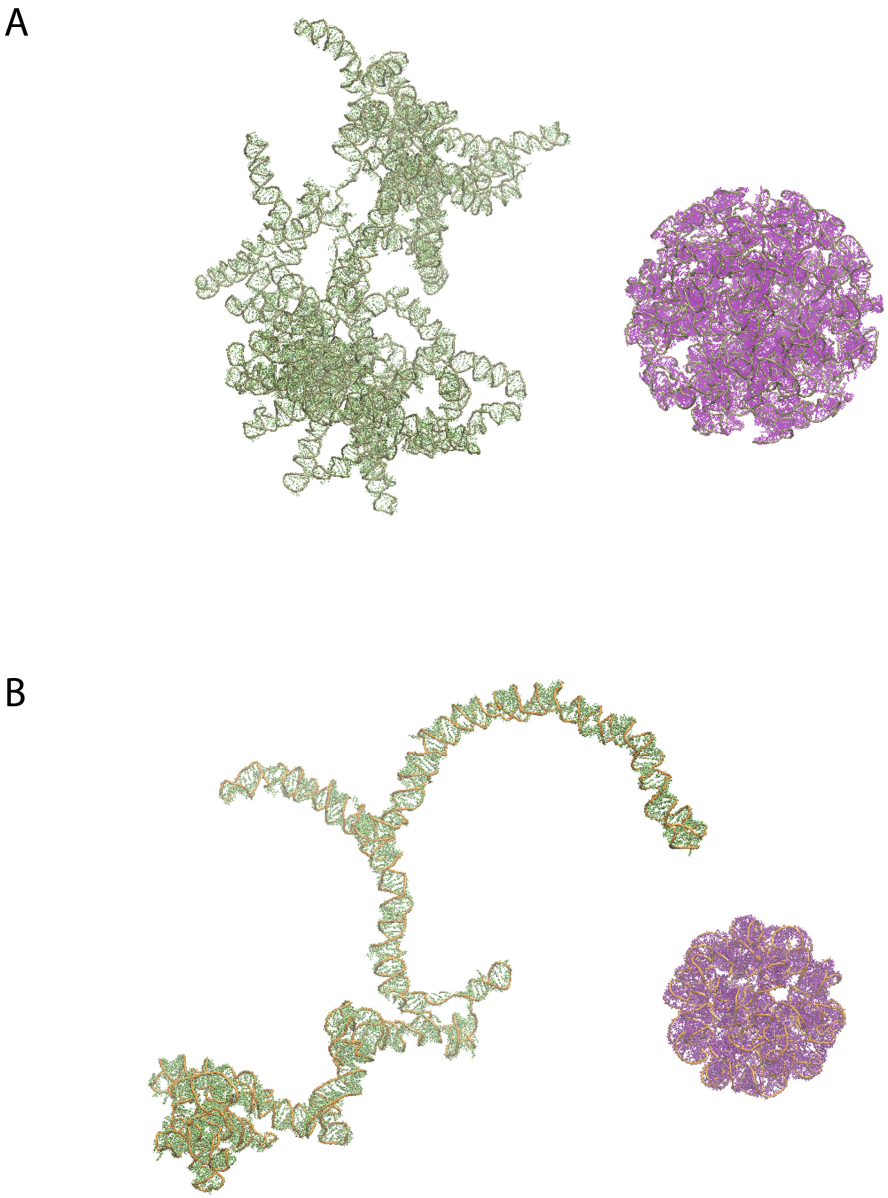

Figure 4.10: Three-dimensional structure of *in vitro* and *in viv*o RNA. A. MS2 RNA (left: *in vitro*; right: *in vivo*) B. STMV RNA (left: *in vitro*; right: *in vivo*)

The compactness of the genomes in PS-positive and PS-negative viruses

A major difference between MS2 and STMV lies in the requirement of a packaging signal. As mentioned earlier, MS2 depends on the TR stem-loop to initiate assembly, while the packaging of STMV RNA does not involve any signal. To examine whether the relationship between packaging signals and RNA compactness is a general feature of viral RNAs, I calculated the MLD for the genomes of different RNA viruses. The results from MLD calculation on MS2 RNA indicate that both the *in vitro* structure (MLD=112) and the structure I predicted for the mature virus (MLD=77) are significantly more compact than the predicted structures of shuffled sequences (<MLD>=346) (Figure 4.11). The MFE structure predicted by RNAfold for the original sequence is also more compact (MLD=160) than the shuffled ones. This suggests that the sequence of MS2 RNA has evolved to enable the formation of a compact three-dimensional structure, which would facilitate the packaging of the genome into the capsid; this is consistent with the proposal of Yoffe *et al.*. Interestingly, the same calculation on STMV leads to completely different results: the predicted *in vivo* structure has an extremely small MLD (MLD=18), while the *in vitro* structure is more extended (MLD=205) than most shuffled sequences (<MLD>=152) (Figure 4.11).

All together the secondary structure analysis suggests that MS2 adopts a different packaging strategy from STMV. It is likely that, because of the requirement of a packaging signal in the middle of the MS2 genome, the RNA has to be synthesized at least half way through before translation of the replicase is suppressed and assembly can be initiated. This probably leads to the evolution of the RNA sequence towards a compact and ready-to-be-packaged structure. As contrast, for the PS-negative STMV, assembly is

likely to proceed simultaneously with RNA replication, during which the protein subunits nucleate the RNA into a series of permanent stem-loops.

Based on the findings, I formed a theory that PS-positive viruses assemble post-replicationally, and the sequences of the viral RNAs have evolved to enable the formation of compact structures that are ready to be packaged. On the other hand, PS-negative viruses assemble co-replicationally, during which the protein subunits interact with the RNA and help it fold into its final structure.

This theory is further supported by MLD calculation on the predicted structures of other viral RNAs. PaV is a T=3 virus that encapsidates two RNAs: RNA1 (3011 nt) and RNA2 (1311 nt) (12). It belongs to the family of nodaviridae, and experiments on several members of nodaviridae have suggested that a packaging signal in form of a RNA stem-loop exist in RNA2 (77). The MLD calculation on PaV RNA2, which has a similar size as STMV RNA, revealed that RNA2 is more compact (MLD=150) than shuffled sequences (<MLD>=170) and much more compact than STMV (Figure 4.11). Unlike what Yoffe *el al.* suggested that the evolution of genomic sequence towards more compact structure is a result of the selection pressure of limited space, my result demonstrated that volume is not always the reason for a more compact viral RNA structure, since the space available for PaV RNA2 is larger than that for STMV RNA. Also, I excluded the possibility that the extendedness of STMV RNA might be a universal phenomenon among T=1 viruses, because I found that STNV, a T=1 virus which is suggested to have an AXXA motif as a packaging signal (78), has a more compact secondary structure (MLD=133) than shuffled sequences (<MLD>=171) (Figure 4.11).

Overall, my results suggest that PS-positive viruses assemble post-replicationally while PS-negative viruses assemble co-replicationally. It would be worthwhile to examine the compactness of other viral RNAs to see if this relationship between packaging signal and RNA compactness is universal among all the viruses.

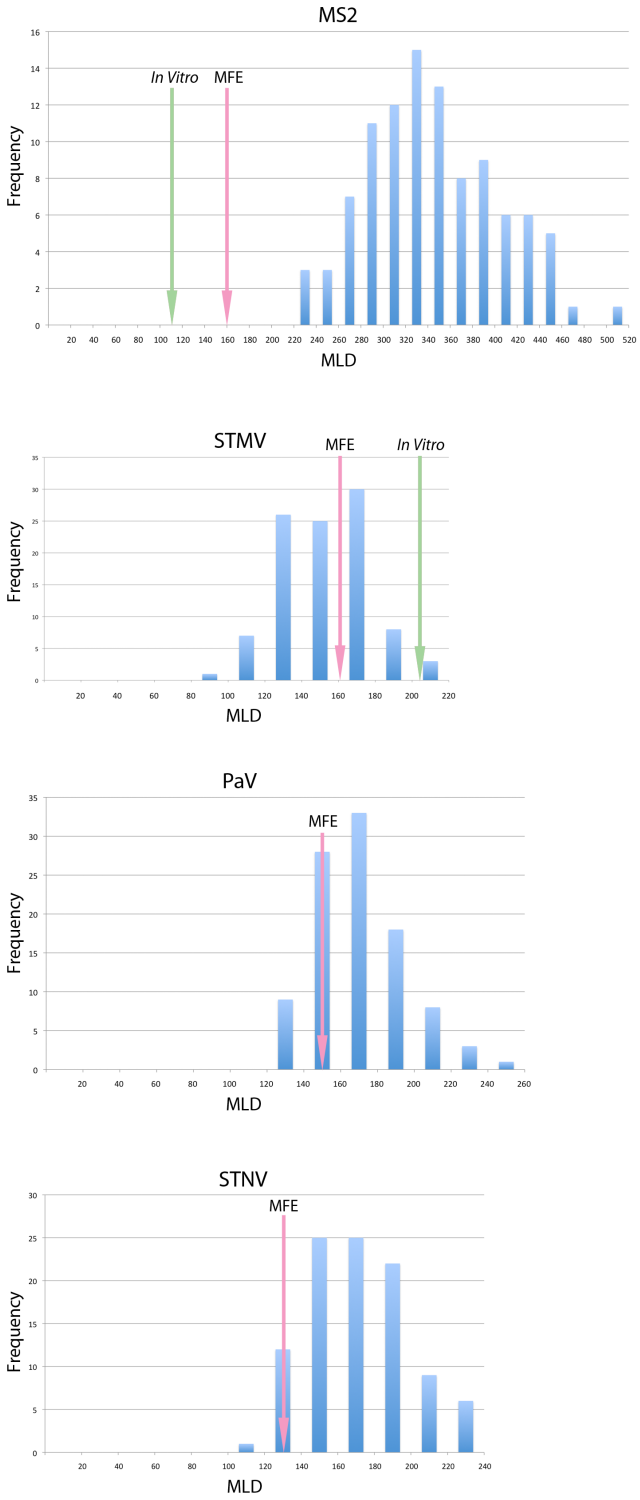

Figure 4.11: MLD of MS2, STMV, PaV and STNV RNAs. (blue columns: structures of shuffled sequences; yellow arrow: MFE structure predicted by RNAfold; green arrow: *in vitro* structure predicted by SELEX and UNAFold (MS2) and *in vitro* structure predicted by SHAPE (STMV))

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

Studying RNA structures in icosahedral single-stranded viruses is important for understanding the assembly of those viral particles. Functions are based on structures, which, in case of viral RNAs, are essential in terms of both stabilization and infection. My studies on viral RNA explored the interrelationships of sequence, structure and function. Starting from RNA secondary structures, I traced back to the RNA sequences to examine the causes of specific structures, and I stepped forward to generate tertiary structural models that reflect their final conformations. The research led to a better vision of how the viral RNAs evolve to form unique structures and how they are packaged into the viral particles.

Evolution of viral RNA sequences

The work on HIV-1 RNA reveals a significantly larger number of adenosines in the genome than that of other RNAs, including 16S and 23S rRNA in different organisms. This composition, together with the sequence of the genome, results in an unusual structure that has a low base pairing frequency. The enriched adenosines are concentrated in the wobble position of the codons, indicating how the virus reconciles the selection pressure on protein coding with the pressure on the formation of a specific form of secondary structure. The fact that adenosines, instead of other nucleotides, are especially abundant suggests that the RNA sequence has evolved to maximize the looseness of the structure, since U or G can pair with C and GC base pairs are more stable than AU base pairs. There might also be tertiary interactions between the adenosines and the proteins.

The large extent of single-strandedness observed in HIV-1 RNA may help the gene replication and translation by facilitating the melting of the helices when the replicase or ribosome is sliding through the RNA. Data from analyses on other viruses suggest that this low base pairing frequency is probably not unique for HIV-1; and it may serve to regulate the tertiary structure besides the roles on replication or translation, as the *in vivo* RNA structures within viral capsids tend to have significantly more unpaired nucleotides than those in the *in vitro* structures of the same RNAs (44,58). My future work will be to examine the sequences and secondary structures of both large and small viruses to see whether this highly unpaired structure is prevalent in all the viruses disregarding the sizes and how the sequences of different viral RNAs evolve to form specific structures.

Modeling of RNA viruses

I have succeeded in modeling the STMV and MS2, both of which recapture the features presented in the crystal structure and cryo-EM density and provide detailed information that is missing in the experimental data. The final model of STMV contains all the protein subunits and the entire genome of 1058 nucleotides. The RMSD between the phosphorus atoms of the model and those of the crystal structure is only 1.26 Å. Comparisons indicate that my model is highly correlated with the crystal structure, and has a better correlation to the X-ray based maps than the McPherson model built from an arbitrary sequence. This model supports the hypothesis that STMV RNA forms thirty local stem-loops during assembly. In addition, an earlier model I built using a UNAFold predicted secondary structure whose stem-loops are mostly asymmetric was proved to

have a rather low correlation with the crystal structure. This fact, together with the success of the final model, suggests that protein subunits are critical in shaping the RNA structure during the assembly process, forcing the RNA into symmetrical stem-loops and interacting with them to stabilize the viral particle.

The MS2 model also provides information on the organization of the genome in three-dimensional space. The model of the RNA has the double-shell structure indicated in cryo-EM density (14), and the predicted secondary structure used for building the model shares a strong similarity with the *in vitro* structure probed in 1980s, indicating that this specific secondary structure is able to adopt a tertiary conformation that is consistent with the experimental images. The models of STMV (T=1) and MS2 (T=3) provide an insight into how the real viral genomic sequences are organized into icosahedral geometries of different scales. The commonality between the two models suggests that viral RNA, when being encapsidated, forms a series of stem-loops or domains that are connected by single-stranded nucleotides, with the stem-loops interacting with the protein capsid and stabilizing the overall geometry while the single-stranded stretches providing the flexibility required for folding the RNA into the target conformation.

So far my modeling focuses on small viruses with a single piece of RNA. In the future, I would like to explore the three-dimensional organizations of viruses with two or more pieces of RNA, such as pariacoto virus and cowpea chlorotic mottle virus. That would shed light on how multiple pieces of RNA are selected and packaged into a single volume and their respective roles in viral assembly. In addition, I would also like to model the structures of larger viruses, especially pathogenic viruses such as hepatitis A virus, SARS

coronavirus and influenza A virus, to see how their structures might be related to the functions, which could help in developing drugs that target those viruses.

The compactness of viral RNA

The modeling and secondary structure studies on RNA viruses indicate a strong correlation between the presence of a packaging signal and the compactness of viral RNA. STMV, which does not require a packaging signal during assembly, has an extended *in vitro* structure with a large MLD; and the large size of the *in vitro* structure in space contrasts with the compact conformation when the same genome is packaged inside the capsid. On the contrary, MS2, whose assembly is initiated by a specific stem-loop as the packaging signal, shows only small differences between the *in vitro* and *in vivo* structures, presenting a relatively small MLD when compared with shuffled sequences. The findings suggest that there are two different assembly pathways existing in viruses: co-replicational assembly for PS-negtive viruses and post-replicational assembly for PS-positive viruses. In PS-negative viruses, since the protein subunits do not have to wait for the formation of a specific stem-loop (the packaging signal), they interact with the RNA immediately as it is replicated to shape it into desired conformations, which costs less energy than refolding all the helices afterwards in order to generate the final structure. On the other hand, in PS-positive viruses, the presence of a specific stem-loop is essential for the accurate packaging of the genome. As a result, the protein subunits have to wait until the signal occurs, and this evolutionary pressure drives the sequence of the RNA towards forming a structure that is compact and ready to be packaged with minimal refolding.

My theory is further supported by MLD calculation on the predicted structures of other viral RNAs. Unlike what Yoffe *el al.* suggested (43), volume is not always the reason for a more compact viral RNA structures. For example, PaV RNA2, which has a similar size as STMV RNA and is provided with a larger space within the viral capsid, is more compact than STMV; and this could be explained by the possibility of having a packaging signal that has been proved to exist in similar viruses of the same family (77). Also, I excluded the possibility that the striking extendedness of STMV RNA is a universal phenomenon among T=1 viruses, since I found that STNV, a T=1 virus which is suggested to have an AXXA motif (78) as a packaging signal, has more compact secondary structures than shuffled sequences. Future research should examine the compactness of other viral RNAs, exploring the possible relationship with the presence or absence of the packaging signal, and to gain more insights into the assembly process of different viruses.

REFERENCES

1. Mirkov, T.E., Mathews, D.M., Duplessis, D.H. and Dodds, J.A. (1989) Nucleotide-Sequence and Translation of Satellite Tobacco Mosaic-Virus Rna. *Virology*, 170, 139-146.

2. Rossmann, M.G. (2002) *Picornavirus structure overview*.

3. Gog, J.R., Afonso, E.D., Dalton, R.M., Leclercq, I., Tiley, L., Elton, D., von Kirchbach, J.C., Naffakh, N., Escriou, N. and Digard, P. (2007) Codon conservation in the influenza A virus genome defines RNA packaging signals. *Nucleic Acids Res*, 35, 1897-1907.

4. Lunel, F. (1992) Hepatitis C virus: the virus responsible of the majority of non A non B hepatitis. 2: Epidemiology of Hepatitis C. *Gastroenterol Clin Biol*, 16, 526-536.

5. Kahn, J.S. (2006) The widening scope of coronaviruses. *Curr Opin Pediatr*, 18, 42-47.

6. Davies, G. (2002) Foot and mouth disease. *Res Vet Sci*, 73, 195-199.

7. Rossmann, M.G. and Johnson, J.E. (1989) Icosahedral RNA virus structure. *Annu Rev Biochem*, 58, 533-573.

8. Caspar, D.L. and Klug, A. (1962) Physical principles in the construction of regular viruses. *Cold Spring Harb Symp Quant Biol*, 27, 1-24.

9. Johnson, J.E. and Speir, J.A. (1997) Quasi-equivalent viruses: a paradigm for protein assemblies. *J Mol Biol*, 269, 665-675.

10. Schneemann, A. (2006) The structural and functional role of RNA in icosahedral virus assembly. *Annu Rev Microbiol*, 60, 51-67.

11. Larson, S.B., Day, J., Greenwood, A. and McPherson, A. (1998) Refined structure of satellite tobacco mosaic virus at 1.8 angstrom resolution. *J Mol Biol*, 277, 37-59.

12. Tang, L., Johnson, K.N., Ball, L.A., Lin, T., Yeager, M. and Johnson, J.E. (2001) The structure of pariacoto virus reveals a dodecahedral cage of duplex RNA. *Nat Struct Biol*, 8, 77-83.

13. Grahn, E., Stonehouse, N.J., Murray, J.B., van den Worm, S., Valegard, K., Fridborg, K., Stockley, P.G. and Liljas, L. (1999) Crystallographic studies of RNA hairpins in complexes with recombinant MS2 capsids: implications for binding requirements. *RNA*, 5, 131-138.

14. Toropova, K., Basnak, G., Twarock, R., Stockley, P.G. and Ranson, N.A. (2008) The three-dimensional structure of genomic RNA in bacteriophage MS2: implications for assembly. *J Mol Biol*, 375, 824-836.

15. Valegard, K., Liljas, L., Fridborg, K. and Unge, T. (1990) The three-dimensional structure of the bacterial virus MS2. *Nature*, 345, 36-41.

16. Stockley, P.G., Stonehouse, N.J., Murray, J.B., Goodman, S.T., Talbot, S.J., Adams, C.J., Liljas, L. and Valegard, K. (1995) Probing sequence-specific RNA recognition by the bacteriophage MS2 coat protein. *Nucleic Acids Res*, 23, 2512-2518.

17. Larson, S.B. and McPherson, A. (2001) Satellite tobacco mosaic virus RNA: structure and implications for assembly. *Curr Opin Struc Biol*, 11, 59-65.

18. Fry, E.E., Grimes, J. and Stuart, D.I. (1999) Virus crystallography. *Mol Biotechnol*, 12, 13-23.

19. Stanley, W.M. (1936) The Inactivation of Crystalline Tobacco-Mosaic Virus Protein. *Science*, 83, 626-627.

20. Klug, A. (1999) The tobacco mosaic virus particle: structure and assembly. *Philos Trans R Soc Lond B Biol Sci*, 354, 531-535.

21. Klug, A. (2010) From virus structure to chromatin: X-ray diffraction to three-dimensional electron microscopy. *Annu Rev Biochem*, 79, 1-35.

22. Abad-Zapatero, C., Abdel-Meguid, S.S., Johnson, J.E., Leslie, A.G., Rayment, I., Rossmann, M.G., Suck, D. and Tsukihara, T. (1980) Structure of southern bean mosaic virus at 2.8 A resolution. *Nature*, 286, 33-39.

23. Chen, Z.G., Stauffacher, C., Li, Y., Schmidt, T., Bomu, W., Kamer, G., Shanks, M., Lomonossoff, G. and Johnson, J.E. (1989) Protein-RNA interactions in an icosahedral virus at 3.0 A resolution. *Science*, 245, 154-159.

24. Fisher, A.J. and Johnson, J.E. (1993) Ordered duplex RNA controls capsid architecture in an icosahedral animal virus. *Nature*, 361, 176-179.

25. Tang, L. and Johnson, J.E. (2002) Structural biology of viruses by the combination of electron cryomicroscopy and X-ray crystallography. *Biochemistry*, 41, 11517-11524.

26. Rossmann, M.G., Morais, M.C., Leiman, P.G. and Zhang, W. (2005) Combining X-ray crystallography and electron microscopy. *Structure*, 13, 355-362.

27. Gan, L., Speir, J.A., Conway, J.F., Lander, G., Cheng, N., Firek, B.A., Hendrix, R.W., Duda, R.L., Liljas, L. and Johnson, J.E. (2006) Capsid conformational sampling in HK97 maturation visualized by X-ray crystallography and cryo-EM. *Structure*, 14, 1655-1665.

28. Lata, R., Conway, J.F., Cheng, N., Duda, R.L., Hendrix, R.W., Wikoff, W.R., Johnson, J.E., Tsuruta, H. and Steven, A.C. (2000) Maturation dynamics of a viral capsid: visualization of transitional intermediate states. *Cell*, 100, 253-263.

29. Toropova, K., Stockley, P.G. and Ranson, N.A. (2011) Visualising a Viral RNA Genome Poised for Release from Its Receptor Complex. *Journal of Molecular Biology*, 408, 408-419.

30. Freddolino, P.L., Arkhipov, A.S., Larson, S.B., McPherson, A. and Schulten, K. (2006) Molecular dynamics simulations of the complete satellite tobacco mosaic virus. *Structure*, 14, 437-449.

31. Devkota, B., Petrov, A.S., Lemieux, S., Boz, M.B., Tang, L., Schneemann, A., Johnson, J.E. and Harvey, S.C. (2009) Structural and electrostatic characterization of pariacoto virus: implications for viral assembly. *Biopolymers*, 91, 530-538.

32. Wuchty, S., Fontana, W., Hofacker, I.L. and Schuster, P. (1999) Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49, 145-165.

33. Hofacker, I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res*, 31, 3429-3431.

34. Markham, N.R. and Zuker, M. (2008) UNAFold: software for nucleic acid folding and hybridization. *Methods Mol Biol*, 453, 3-31.

35. Swenson, M.S., Anderson, J., Ash, A., Gaurav, P., Sukosd, Z., Bader, D.A., Harvey, S.C. and Heitsch, C.E. (2012) GTfold: Enabling parallel RNA secondary structure prediction on multi-core desktops. *BMC Res Notes*, 5, 341.

36. Xia, T., SantaLucia, J., Jr., Burkard, M.E., Kierzek, R., Schroeder, S.J., Jiao, X., Cox, C. and Turner, D.H. (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry*, 37, 14719-14735.

37. Freier, S.M., Kierzek, R., Jaeger, J.A., Sugimoto, N., Caruthers, M.H., Neilson, T. and Turner, D.H. (1986) Improved free-energy parameters for predictions of RNA duplex stability. *Proc Natl Acad Sci U S A*, 83, 9373-9377.

38. McCaskill, J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29, 1105-1119.

39. Deigan, K.E., Li, T.W., Mathews, D.H. and Weeks, K.M. (2009) Accurate SHAPE-directed RNA structure determination. *Proc Natl Acad Sci U S A*, 106, 97-102.

40. Vasa, S.M., Guex, N., Wilkinson, K.A., Weeks, K.M. and Giddings, M.C. (2008) ShapeFinder: a software system for high-throughput quantitative analysis of nucleic acid reactivity information resolved by capillary electrophoresis. *RNA*, 14, 1979-1990.

41. Mathews, D.H. and Turner, D.H. (2002) Experimentally derived nearest-neighbor parameters for the stability of RNA three- and four-way multibranch loops. *Biochemistry*, 41, 869-880.

42. Watts, J.M., Dang, K.K., Gorelick, R.J., Leonard, C.W., Bess, J.W., Jr., Swanstrom, R., Burch, C.L. and Weeks, K.M. (2009) Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature*, 460, 711-716.

43. Yoffe, A.M., Prinsen, P., Gopal, A., Knobler, C.M., Gelbart, W.M. and Ben-Shaul, A. (2008) Predicting the sizes of large RNA molecules. *Proc Natl Acad Sci U S A*, 105, 16153-16158.

44. Athavale, S.S., Gossett, J.J., Bowman, J.C., Hud, N.V., Williams, L.D. and Harvey, S.C. (2013) In vitro secondary structure of the genomic RNA of satellite tobacco mosaic virus. *PLoS One*, 8, e54384.

45. Gutell, R.R., Power, A., Hertz, G.Z., Putz, E.J. and Stormo, G.D. (1992) Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucleic Acids Res*, 20, 5785-5795.

46. Gutell, R.R., Lee, J.C. and Cannone, J.J. (2002) The accuracy of ribosomal RNA comparative structure models. *Curr Opin Struct Biol*, 12, 301-310.

47. Noller, H.F. and Woese, C.R. (1981) Secondary structure of 16S ribosomal RNA. *Science*, 212, 403-411.

48. Zuker, M. (1989) On finding all suboptimal foldings of an RNA molecule. *Science*, 244, 48-52.

49. Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*, 31, 3406-3415.

50. Mathews, D.H., Sabina, J., Zuker, M. and Turner, D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol*, 288, 911-940.

51. Schuster, P. (2006) Prediction of RNA secondary structures: from theory to models and real molecules. *Rep. Prog. Phys.*, 59.

52. Merino, E.J., Wilkinson, K.A., Coughlan, J.L. and Weeks, K.M. (2005) RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). *J Am Chem Soc*, 127, 4223-4231.

53. Mortimer, S.A. and Weeks, K.M. (2007) A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. *J Am Chem Soc*, 129, 4144-4145.

54. Zuker, M. and Stiegler, P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*, 9, 133-148.

55. Larson, S.B., Koszelak, S., Day, J., Greenwood, A., Dodds, J.A. and Mcpherson, A. (1993) 3-Dimensional Structure of Satellite Tobacco Mosaic-Virus at 2.9 Angstrom Resolution. *J Mol Biol*, 231, 375-391.

56. Larson, S.B., Koszelak, S., Day, J., Greenwood, A., Dodds, J.A. and Mcpherson, A. (1993) Double-Helical Rna in Satellite Tobacco Mosaic-Virus. *Nature*, 361, 179-182.

57. Kuznetsov, Y.G., Dowell, J.J., Gavira, J.A., Ng, J.D. and McPherson, A. (2010) Biophysical and atomic force microscopy characterization of the RNA from satellite tobacco mosaic virus. *Nucleic Acids Res*, 38, 8284-8294.

58. Schroeder, S.J., Stone, J.W., Bleckley, S., Gibbons, T. and Mathews, D.M. (2011) Ensemble of secondary structures for encapsidated satellite tobacco mosaic virus RNA consistent with chemical probing and crystallography constraints. *Biophys J*, 101, 167-175.

59. Parisien, M. and Major, F. (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, 452, 51-55.

60. Phillips, J.C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R.D., Kale, L. and Schulten, K. (2005) Scalable molecular dynamics with NAMD. *J Comput Chem*, 26, 1781-1802.

61. Humphrey, W., Dalke, A. and Schulten, K. (1996) VMD: Visual Molecular Dynamics. *J. Mol. Graphics*, 14, 33-38.

62. Jonikas, M.A., Radmer, R.J., Laederach, A., Das, R., Pearlman, S., Herschlag, D. and Altman, R.B. (2009) Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA*, 15, 189-199.

63. Jonikas, M.A., Radmer, R.J. and Altman, R.B. (2009) Knowledge-based instantiation of full atomic detail into coarse-grain RNA 3D structural models. *Bioinformatics*, 25, 3259-3266.

64. Larson, S.B., Day, J., Greenwood, A. and McPherson, A. (1998) Refined structure of satellite tobacco mosaic virus at 1.8Å resolution. *J Mol Biol*, 277, 37-59.

65. Dreher, T.W. (2009) Role of tRNA-like structures in controlling plant virus replication. *Virus research*, 139, 217-229.

66. Felden, B., Florentz, C., McPherson, A. and Giege, R. (1994) A histidine accepting tRNA-like fold at the 3'-end of satellite tobacco mosaic virus RNA. *Nucleic Acids Res*, 22, 2882-2886.

67. Reeder, J., Steffen, P. and Giegerich, R. (2007) pknotsRG: RNA pseudoknot folding including near-optimal structures and sliding windows. *Nucleic Acids Res*, 35, W320-324.

68. Rodriguez-Alvarado, G. and Roossinck, M.J. (1997) Structural analysis of a necrogenic strain of cucumber mosaic cucumovirus satellite RNA in planta. *Virology*, 236, 155-166.

69. Kozak, M. and Nathans, D. (1971) Fate of Maturation Protein during Infection by Coliphage Ms2. *Nature-New Biol*, 234, 209-&.

70. Choi, Y.G. and Rao, A.L.N. (2003) Packaging of brome mosaic virus RNA3 is mediated through a bipartite signal. *J Virol*, 77, 9750-9757.

71. Kawamoto, S., Ueda, K., Mita, E. and Matsubara, K. (1994) The Packaging Signal in Hepatitis-B Virus Pregenome Functions Only at the 5' End. *J Virol Methods*, 49, 113-127.

72. Annamalai, P. and Rao, A.L.N. (2005) Dispensability of 3' tRNA-like sequence for packaging cowpea chlorotic mottle virus genomic RNAs. *Virology*, 332, 650-658.

73. Dykeman, E.C., Grayson, N.E., Toropova, K., Ranson, N.A., Stockley, P.G. and Twarock, R. (2011) Simple rules for efficient assembly predict the layout of a packaged viral RNA. *J Mol Biol*, 408, 399-407.

74. Zeng, Y., Larson, S.B., Heitsch, C.E., McPherson, A. and Harvey, S.C. (2012) A model for the structure of satellite tobacco mosaic virus. *J Struct Biol*, 180, 110-116.

75. Popenda, M., Szachniuk, M., Antczak, M., Purzycka, K.J., Lukasiak, P., Bartol, N., Blazewicz, J. and Adamiak, R.W. (2012) Automated 3D structure composition for large RNAs. *Nucleic Acids Res*, 40, e112.

76. Zhang, Y., Arakaki, A.K. and Skolnick, J. (2005) TASSER: an automated method for the prediction of protein tertiary structures in CASP6. *Proteins*, 61 Suppl 7, 91-98.

77. Zhong, W.D., Dasgupta, R. and Rueckert, R. (1992) Evidence That the Packaging Signal for Nodaviral Rna2 Is a Bulged Stem Loop. *P Natl Acad Sci USA*, 89, 11146-11150.

78. Bunka, D.H.J., Lane, S.W., Lane, C.L., Dykeman, E.C., Ford, R.J., Barker, A.M., Twarock, R., Phillips, S.E.V. and Stockley, P.G. (2011) Degenerate RNA Packaging Signals in the Genome of Satellite Tobacco Necrosis Virus: Implications for the Assembly of a T=1 Capsid. *J Mol Biol*, 413, 51-65.